University of California
Santa Barbara

Enabling next generation mobile communication via millimeter-wave technology

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Electrical and Computer Engineering

by

Maryam Eslami Rasekh

Committee in charge:

Professor Upamanyu Madhow, Chair
Professor Dongning Guo, Northwestern University
Professor Joao Hespanha
Professor Ramtin Pedarsani

March 2020

The Dissertation of Maryam Eslami Rasekh is approved.

Professor Dongning Guo, Northwestern University

Professor Joao Hespanha

Professor Ramtin Pedarsani

Professor Upamanyu Madhow, Committee Chair

September 2019

Enabling next generation mobile communication via millimeter-wave technology

by

Maryam Eslami Rasekh

To Maman, Baba, and Marzie,
your place was empty beside me.

Acknowledgements

My gratitude goes to so many people who have helped and supported me over the past few years. First and foremost, I am grateful to my wonderful advisor, Professor Upamanyu Madhow, who has been a true mentor for me, academically and otherwise. Under his supervision, I gained valuable expertise in many areas that were new to me, and came into contact with amazing people and projects in academia and industry. His support has been unwavering through all the challenges of my time here, and I consider myself very lucky to have joined his group. I would also like to thank my committee members, Prof. Dongning Guo, Prof. Joao Hespanha, and Prof. Ramtin Pedarsani for their insightful comments from which I have learned a lot.

I am extremely grateful for my collaborations with Prof. Dongning Guo, Prof. Mark Rodwell, and Prof. Heather Zheng. Working with them has been an experience that I cherish and from which I have gained immeasurably. I would also like to thank my collaborators in the WCSL group, Mohammed Abdelghany, Faruk Gencel and Zhinus Marzi for their insightful discussions. It has been a true pleasure working with them, alongside our colleagues in the lab, Ahmet, Anant, Andrew, Aseem, Babak, Can, Dinesh, Francois, Metehan, Soorya and Sriram who were the best labmates one could wish for. My time in Santa Barbara has been a delight thanks to my wonderful labmates, amazing professors and beloved friends.

Above all, I owe everything to my amazing parents, Pouran and Abbas, who were the ultimate teachers of my life, my sister, Marzie, and my entire extended family. It truly does take a village and I am infinitely grateful for mine.

My research was supported by NSF grants CNS-1518812 and CNS-1317153, and by ComSenTer, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

Curriculum Vitæ

Maryam Eslami Rasekh

Education

2020	Doctor of Philosophy, Electrical and Computer Engineering University of California Santa Barbara, USA.
2009	Master of Science, Electrical Engineering/Communications and RF Sharif University of Technology, Tehran, Iran.
2007	Bachelor of Science, Electrical Engineering/Communications Isfahan University of Technology, Isfahan, Iran.

Publications

J4 **Maryam Eslami Rasekh**, Mohammed Abdelghany, Upamanyu Madhow and Mark Rodwell, *Phase noise in modular architectures for millimeter wave massive MIMO,* under review, arXiv preprint arXiv:1910.09095

J3 **Maryam Eslami Rasekh**, Dongning Guo and Upamanyu Madhow, *Joint routing and resource allocation for millimeter wave picocellular backhaul,* to appear in IEEE Transactions on Wireless Communications, 2019

C11 **Maryam Eslami Rasekh**, Mohammed Abdelghany, Upamanyu Madhow and Mark Rodwell, *Phase noise analysis for mmwave massive MIMO: A design framework for scaling via tiled architectures,* 53rd Annual Conference on Information Sciences and Systems (CISS), pp. 1-6. IEEE, 2019.

C10 **Maryam Eslami Rasekh** and Upamanyu Madhow, *Noncoherent compressive channel estimation for mm-wave massive MIMO,* 52nd Asilomar Conference on Signals, Systems, and Computers, pp. 889-894. IEEE, 2018.

J2 Mohammed Faruk Gencel, **Maryam Eslami Rasekh** and Upamanyu Madhow, *Distributed base station: A concept system for long-range broadband wireless access,* arXiv preprint arXiv:1803.02877 (2018).

C9 **Maryam Eslami Rasekh**, Zhinus Marzi, Yanzi Zhu, Upamanyu Madhow and Heather Zheng, *Noncoherent mmWave path tracking,* Proceedings of the 18th International Workshop on Mobile Computing Systems and Applications, pp. 13-18. ACM, 2017.

C8 **Maryam Eslami Rasekh**, Dongning Guo and Upamanyu Madhow, *Interference-aware routing and spectrum allocation for millimeter wave backhaul in urban picocells,* 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 1-7. IEEE, 2015.

C7 Mohammed Faruk Gencel, **Maryam Eslami Rasekh** and Upamanyu Madhow, *Noise-resilient scaling for wideband distributed beamforming,* 49th Asilomar Conference on Signals, Systems and Computers, pp. 276-280. IEEE, 2015.

C6 Mohammed Faruk Gencel, **Maryam Eslami Rasekh** and Upamanyu Madhow, *Distributed transmit beamforming with one bit feedback revisited: How noise limits scaling,* International Symposium on Information Theory (ISIT). IEEE, 2015.

C5 Mohammed Faruk Gencel, **Maryam Eslami Rasekh** and Upamanyu Madhow, *Scaling wideband distributed transmit beamforming via aggregate feedback,* International Conference on Communications (ICC), pp. 2356-2362. IEEE, 2015.

C4 **Maryam Eslami Rasekh**, Raghuraman Mudumbai and Upamanyu Madhow, *Frequency tracking with intermittent wrapped phase measurement using the Rao-Blackwellized particle filter,* 48th Asilomar Conference on Signals, Systems and Computers. IEEE, 2014.

J1 **Maryam Eslami Rasekh** and Forouhar Farzaneh, *Interference analysis in an urban mesh network operating in the 60GHz band,* ETRI Journal 35, no. 5 (2013): 775-785.

C3 **Maryam Eslami Rasekh** and Forouhar Farzaneh, *Interference control using polarization in directive 60 GHz mesh networks,* Second Conference on Millimeter-Wave and Terahertz Technologies (MMWaTT), pp. 13-16. IEEE, 2012.

C2 **Maryam Eslami Rasekh**, Forouhar Farzaneh and Amir Ahmad Shishegar, *A street canyon approximation model for the 60 GHz propagation channel in an urban environment with rough surfaces,* 5th International Symposium on Telecommunications, pp. 132-137. IEEE, 2010.

C1 **Maryam Eslami Rasekh**, Amir Ahmad Shishegar and Forouhar Farzaneh, *A study of the effect of diffraction and rough surface scattering modeling on ray tracing results in an urban environment at 60 GHz,* First Conference on Millimeter-Wave and Terahertz Technologies (MMWaTT), pp. 27-31. IEEE, 2009.

Abstract

Enabling next generation mobile communication via millimeter-wave technology

by

Maryam Eslami Rasekh

To adapt with current trends in wireless data consumption, providers need to increase capacity by orders of magnitude. Millimeter (mm) wave communication has the potential for delivering this boost through unprecedented levels of spatial reuse. High non-line-of-sight propagation loss at small wavelengths produces a highly localized interference pattern, enabling dense deployment of mobile access points each serving tiny "picocells", 10's of meters in diameter. On the other hand, small wavelengths allow packing massive antenna arrays in reasonable form factors that can be mounted on street fixtures such as lampposts, and are able to form very narrow beams that provide significant spatial isolation. We first consider the problem of **backhaul support** for the dense deployment of picocellular access points. As wired backhauling is not scalable, we propose a multihop mesh network of high-bandwidth directional point-to-point links at mmWave or THz frequencies to connect picocellular access points to the existing wired gateways, and develop a framework for joint optimization of routing and link scheduling for maximizing backhaul throughput. Scaling array sizes, on the other hand, poses serious challenges both in signal processing and in hardware implementation. As arrays grow large, their beams becomes narrow and accurate **user tracking** is required to maintain a beamformed link. Current mmWave hardware offers only analog RF beamforming capacity, and most systems do not maintain phase coherence from one packet to another. We therefore propose a noncoherent compressive channel estimation framework with logarithmic overhead scaling that relies solely on magnitude measurements. As technology

matures, the realization of *fully digital* beamforming with one RF chain per antenna element will enable the synthesis of multiple simultaneous beams on a single frontend. Hardware impairments are often more serious for wideband communication at higher frequencies, and a systematic framework for quantifying their impact is crucial. We develop such a framework for analyzing the effect of **phase noise** on mmWave multi-user massive MIMO, scaled using a tiled architecture. Our analysis provides a cross-layer design tool which can be employed by hardware designers to determine allowable masks for the phase noise power spectral density for different circuit components.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As data-hungry mobile devices and applications such as live streaming (both in downlink and uplink) become increasingly prevalent, the mobile communication infrastructure needs to evolve dramatically to support the exploding demand for wireless data. The shift to higher frequency bands is already under way with significant efforts to develop and implement 5G, and will likely continue to colonize even higher frequencies well into the terahertz (THz) spectrum. Of course, bandwidth alone will not meet the mobile demand of a large number of users; bridging this gap in the long term is only possible with aggressive frequency reuse.

Small cells, large arrays. There are two main approaches to increasing frequency reuse. The first is reducing cell sizes so that fewer users share the resources of one base station (or access point), and the second is spatial multiplexing via multiple antenna systems. Luckily, both of these avenues are strong points for mmWave communication. Small wavelengths incur higher free-space propagation loss and, on top of that, atmospheric absorption losses. These factors limit viable communication to short line-of-sight (LoS) links with distances of 10s or at most 100s of meters. Consequently, interference is limited locally, making it possible to reuse the same frequency band many times in a

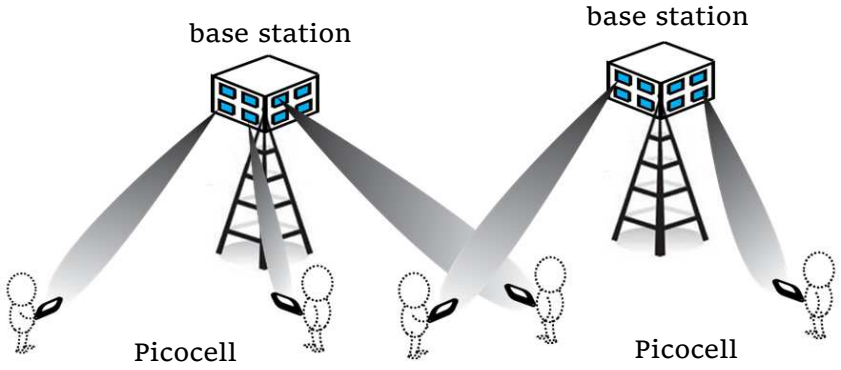


Figure 1.1: Closely-spaced picocellular base stations with directional antennas serving many users simultaneously.

small area. Furthermore, even at small distances, mmWave links require beamforming gain to establish high capacity communication. At small wavelengths, massive arrays with 100s or even 1000s of elements can fit in small form factors with half-wavelength spacing. These highly directional yet compact arrays can be deployed opportunistically on lampposts or walls as "picocellular" access points capable of supporting many users simultaneously through spatial multiplexing. In this thesis, we address some of the challenges in realizing smaller cells and directional communication.

1.1 Backhaul support for picocellular networks

Picocellular architectures comprised of closely-spaced access points with intense spatial reuse play a critical role in the evolution of mobile systems, particularly in high-density urban and suburban environments [1]. LTE data rates are projected to approach gigabits per second peak rates through carrier aggregation, which is likely to be extended further in next generation networks by using 60 GHz (or other unlicensed mmWave bands) directly from pico base station to the mobile [2], assuming that significant challenges due to blockage and mobility can be overcome. Indeed, a recent interference analysis for such networks [3] indicates that capacities of the order of terabits per second

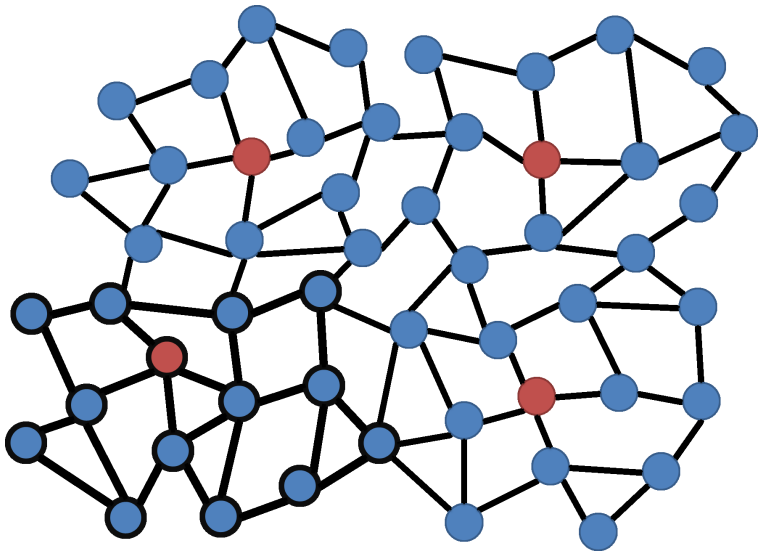


Figure 1.2: Extending backhaul form wired gateways (red nodes) to the dense grid of picocellular access points via a wireless mesh with directional links.

per kilometer (along a single urban canyon) can be obtained with only a few GHz of spectrum by taking advantage of the aggressive spatial reuse enabled by highly directional mmWave links. Moreover, this capacity roughly adds up across parallel canyons, given the strong isolation provided by building blockage.

Delivering such high data rates to mobile users requires that the pico base stations have a sufficiently high-capacity backhaul connection to the Internet. For opportunistic picocellular deployments on lampposts and rooftops, providing optical fiber connectivity for each base station is practically impossible in the foreseeable future, so that wireless backhaul becomes a natural choice [4–6].

In Chapter 2, we consider a mesh network with highly directional mmWave links as a means of extending backhaul from wired gateways to the picocell access points (Figure 1.1). Each access point is a node in the wireless mesh network and is connected to neighboring nodes through high-speed directional mmWave links. The objective is to

route traffic between base stations and gateways through multihop paths, such that each picocell can support a given level of downlink and uplink throughput on the access links to mobile devices. We assume that directional antennas are used on each backhaul link, but because of half duplex communication and residual interference, these links cannot be treated as wires. In particular, in long urban canyons where even non-contiguous links are likely to be aligned, the interference between distant links must be taken into account in deriving the optimal resource allocation and routing.

1.1.1 Contributions

We formulate the wireless mesh backhaul design problem as a joint routing and resource allocation optimization, with the goal of maximizing the access rate at the base stations, while accounting for the mutual interference between simultaneously active links. Our framework applies to any mesh backhaul network via the following abstraction: For any set of simultaneously active links, we must be able to compute the achievable data rate on each active link while accounting for the interference from other active links. However, the computational complexity depends on the nature of the interference patterns, which depends on the propagation environment, the antenna patterns, and the carrier frequency. Specifically, the highly directive nature of mmWave links leads to localized and sparse interference that we are able to exploit for efficient computation of the optimal solution. We illustrate our approach for two settings, urban and suburban, which exhibit interference patterns that are quite different. In built-up urban environments, the highly directional nature of the links, and the ease of blockage of mmWaves by buildings, imply that each link incurs interference only with links within the same street. However, the interference among such links can be significant due to their similar alignment. In more open suburban environments, on the other hand, a larger number of

links can cause interference on each other, but links are less likely to be aligned, so that interference is typically severely attenuated by antenna directionality at the transmitter and the receiver.

Key contributions of the proposed architecture and analysis include the following:

- *Accurate interference modeling:* Environment-specific propagation models are used to derive a realistic interference graph for the network and the optimal level of spatial reuse is maintained, yielding the highest possible backhaul throughput upon deployment. The solutions proposed here can be easily generalized to a variety of network structures with different antenna patterns and arbitrary interference models.

- *Arbitrary (fixed-rate) traffic flow:* We first develop a framework for *downlink* traffic routing and resource allocation. We then extend the formulation to perform joint uplink and downlink optimization. Additional flows between nodes in the network can be included in the optimization by introducing a new commodity for each flow. While we only consider fixed-rate traffic flows, extension to any concave objective of the rate vector is possible in the proposed framework.

- *Arbitrary topologies:* The proposed optimization framework applies to arbitrary network structures with limited node degrees. The networks in our simulations are comprised of several gateways that are connected to all nodes through multihop paths. Each access point is connected to all neighbors that are close enough to form a connection (the number of neighbors is not large in practice). The existence of various paths from each pico base station to different gateways provides redundancy in network resources and improves backhaul reliability, allowing the network to adapt to disruption of links due to blockage or hardware failure. Backhaul in future cellular networks may constitute a mix of wireless *and wired* links. Wired

links can easily be incorporated in the proposed framework as links with capacity constraint that neither cause nor are affected by interference.

- *Versatile and efficient optimization framework:* The problem of resource allocation for optimizing backhaul throughput is first formulated as a linear program, the dimensions of which grow exponentially with network size. Demonstrating the existence of a sparse solution to this linear program, we formulate an equivalent mixed-integer linear program that scales linearly with network size. The proposed formulation is able to solve relatively large networks with hundreds of nodes in a short period of time.

We first present a combinatorial formulation of the optimal resource allocation and routing problem where resources are divided between all possible activation patterns, or *subsets* of simultaneously active links [7]. This blows up the problem size exponentially, going from a scheduling of L links to 2^L possible link combinations, but enables formulation of backhaul scheduling as a linear program. Solving this problem is straightforward for small networks but quickly becomes intractable for larger graphs.[1] One important result, guaranteed by Caratheodory's theorem [8], is that a sparse solution exists for the combinatorial formulation with at most N active patterns, where N is the number of nodes in the network. We thus provide a scalable optimization framework which exploits the existence of an N-sparse solution, by reformulating the problem in terms of *local* interference parameters [9]. The result is a binary linear program (BLP) that scales near-linearly with network size, and can be solved relatively quickly using branch and bound techniques for larger networks with multiple gateways and hundreds of nodes. We also extend the formulation to incorporate joint scheduling for uplink and downlink traffic, or any other proportionally defined set of traffic flows.

[1] One suboptimal approach is to *cluster* the network around gateways, as shown in Figure 1.1, and solve each cluster independently.

1.2 Low-overhead user tracking

Small wavelengths allow scaling to massive MIMO with very large number of elements on compact base stations deployable, for example, on lampposts or other street structures. The narrow pencil beams formed by such arrays enable drastically increased spatial reuse compared to existing cellular networks. A critical challenge in realizing such links, however, is agile adaptation of large arrays to track mobile users while accounting for frequently occurring blockages.

As array sizes grow, simplification of front-ends is critical. Existing mmWave transceivers employ RF beamforming in place of per-element baseband control. This means a single I/Q stream is upconverted at the transmitter and distributed to all array elements, and the phase of each element is tuned with RF phase shifters. At the receiver, the phase of each element is manipulated at RF and the combination of element outputs provides a single I/Q stream for signal processing. Thus, classical least squares techniques, which require individual access to the I/Q signal at each antenna element, cannot be applied for adaptive beamforming.

Conventional techniques for discovering spatial paths when constrained to RF beamforming include exhaustive and hierarchical scan. In exhaustive scan, the transmitter scans the entire angular domain with its narrow beam to identify the strongest path(s) to the user. The number of measurements scales linearly with array size, hence the overhead becomes prohibitive for large arrays. In hierarchical scan, the entire angular space is initially scanned with a small number of broad beams, with feedback from the receiver used to successively narrow the search space. The number of measurements scales logarithmically with array size, but waiting for feedback from the receiver before each scan can be highly time-consuming and impractical upon implementation. This method does not scale well with the number of users, since each user may require a different beacon

sequence, depending on its location and feedback. Neither of these methods is therefore suitable for low-overhead tracking for mmWave massive MIMO.

A promising alternative is to employ *compressive* techniques, introduced in [10, 11] and discussed in detail in the context of picocellular networks in [12]. These techniques leverage the inherent sparsity of the mmWave channel to track users with a small number of measurements. In this approach, the transmitter broadcasts beacons using pseudorandom phases, and uses feedback from the receiver regarding the complex gain observed for each beacon to estimate the spatial channel by identifying the dominant paths. Such schemes can be implemented using RF beamforming (i.e., a single RF chain, rather than one RF chain per element) with severely quantized phase control, which allows simplification of the RF front end.

We term the approach in [10–12] *coherent* compressive estimation, since the receiver must maintain phase coherence across successive measurements in order to provide the desired complex gains as feedback to the transmitter. Unfortunately, such an approach does not work with commodity hardware, since current mmWave systems such as the 802.11ad standard are not designed to maintain phase coherence across packets, and the oscillator offset and drift between the transmitter and receiver can alter the phase of each channel measurement randomly. This motivates us to develop the *noncoherent* compressive estimation approach presented in Chapter 3.

1.2.1 Contributions

We use the same compressive beaconing strategy as in prior work [10–12], but provide an algorithm that can estimate a sparse spatial channel from RSS measurements alone. It does not require phase coherence across beacons, and can therefore be realized with commodity hardware.

Recovery of the single strongest path. We first consider single-path channels with one dominant path between base station and user. In this case we propose noncoherent template matching on an oversampled grid followed by Newton refinement of the grid point with highest correlation with the RSS measurements [13]. Using a 60GHz phased array testbed provided by the Facebook Terragraph team [14], we performed experimental validation of this method. Each testbed radio operates under the 802.11ad standard, and has a 8×16 phased array (128 elements), with effectively 16 horizontally steerable elements (each consisting of 8 vertical elements with fixed phase relationships). Experimental results show that our noncoherent design achieves accuracy similar to that of exhaustive beam scanning but with significantly lower overhead. This advantage increases sharply with the number of steerable elements, as demonstrated by simulations.

Multi-path channel estimation. For recovering multiple viable paths in the channel, we propose and evaluate a two-stage algorithm [15]. The first is phase retrieval, in which the complex value of compressive measurements is recovered up to a constant phase offset, followed by coherent compressive estimation using the output of the phase retrieval stage. We show that the number of measurements for sparse channel recovery with noncoherent compressive estimation scales only slightly worse than for coherent estimation.

The two-stage approach requires decomposition of the measurement matrix as the product of a phase retrieval matrix and a compressive measurement matrix. It is known that matrices with independent and identically distributed (i.i.d.) complex Gaussian entries are effective for this purpose, but the product of such matrices has entries with complex values lying in a continuum, which cannot be realized with coarse phase control. In our proposed approach, we constrain the *product* of the two matrices (i.e., the actual measurement matrix) to be implementable with coarse phase control, and decompose it into two virtual matrices: an inner matrix that is used for coherent compressive estimation, and an outer matrix that is used for phase retrieval. That is, we choose one of the

matrices, and infer the other via a pseudoinverse to minimize the quantization error. We provide design guidelines for the (non-unique) virtual decomposition, and demonstrate its performance and scalability through simulations.

1.3 Multi-user MIMO in phase noise

Millimeter wave (mmWave) bands offer potential channel bandwidths of several GHz. While there have been significant advances in low-cost mmWave hardware over the past decade, existing mmWave hardware is typically based on RF beamforming, which enables synthesis of only one beam at a time. However, we expect fully digital beamforming to become feasible in the relatively near future, with one RF chain per antenna element: this will enable the synthesis of multiple simultaneous beams, and lead to truly massive MIMO with 100s or 1000s of antenna elements, supporting 100s of simultaneous users, each at rates of Gbps. Hardware impairments can be substantial at such high frequencies, especially as we scale to massive arrays and large bandwidth. In Chapter 4, we analyze the impact of one potential bottleneck in such a system: phase noise. A systematic framework for determining the impact of phase noise on MIMO performance is important because oscillators at higher carrier frequencies typically exhibit higher noise power spectral density (PSD), and the increased system bandwidth that we must integrate over contributes further to increased distortion in the phase.

In order to scale to a large number of elements, we consider a modular approach in which a large array is built from subarrays, or tiles, with separate RF processing on each tile. We focus on uplink massive MIMO, hence the RF processing in a tile consists of downconversion to baseband, followed by quantization. Thus, only digital baseband signals are communicated from each tile to a central processor. While this approach greatly simplifies RF hardware design (each tile can be controlled by a separate RF chip)

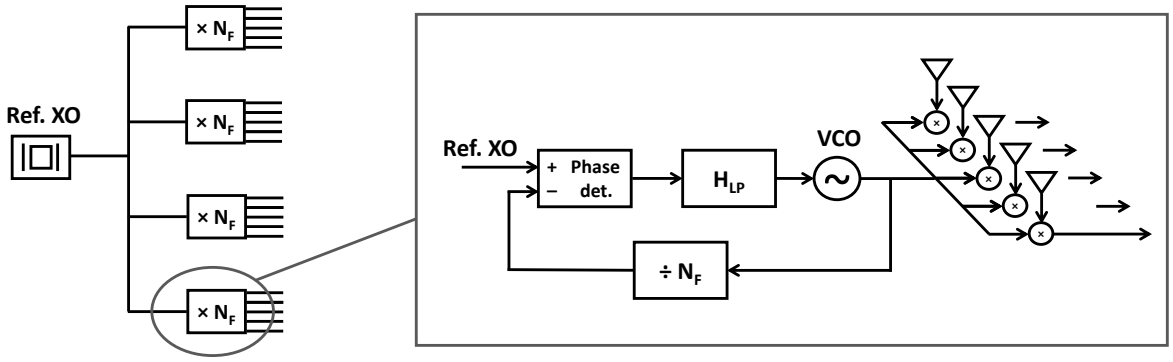


Figure 1.3: Two-step clock distribution in the tiled array.

and produces an easily expandable system, synchronizing the subarrays with each other requires distribution of a common clock. Distributing a high-frequency reference over a large area incurs high propagation loss and is therefore inefficient. We consider here an alternative approach, in which a lower frequency clock is distributed to the tiles, with each tile synthesizing its local oscillator using, for example, a harmonic multiplier or a phase locked loop, as shown in Figure 1.3.

1.3.1 Contributions

We analyze the effect of phase noise on a tiled architecture, where phase noise is generated both in the reference oscillator and in independent oscillators that are locked to the reference on each tile [16, 17]. We first derive the optimal LMMSE receiver for multi-user detection in the presence of phase noise, and proceed to assess the system-level impact of phase noise. For this purpose, we consider a MIMO uplink with K simultaneous users communicating with a base station using a tiled phased array, with singlecarrier modulation and linear MMSE (LMMSE) reception. While our approach applies to a variety of frequency multiplication techniques, for concreteness, we focus on PLL-based generation of the local oscillator in each tile. We model propagation of

phase noise through the two-step clock generation using a linearized PLL model. Our conclusions are summarized as follows:

(a) The common phase noise arising from the low-frequency clock "passes through" the LMMSE receiver, and contributes to symbol rotation after interference suppression. Since the PLL in each tile acts as a lowpass filter, the phase rotations across symbols vary slowly and can be tracked effectively. Thus, for typical settings, the impact of common phase noise is small.

(b) The independent phase noise from the voltage controlled oscillator (VCO) in each tile sees a highpass response from the PLL, and cannot be tracked at the symbol time scale. If LMMSE reception is applied for a nominal set of channels, the effect of phase noise is to create residual multiuser interference. We provide an approximate analysis which shows that, for typical settings, we get substantial benefit by averaging across tiles. Simulation results are provided to demonstrate these predictions.

(c) The dependence of the phase noise PSD on frequency is often modeled as $L(f) = a_0 + \frac{a_1}{f} + \frac{a_2}{f^2} + \frac{a_3}{f^3}$, with parameters $\{a_i\}$ that depend on the oscillator structure and design. We note that, under our linearlized PLL model, these parameters can be directly related to the output phase noise variances. We use this observation to translate output performance criteria such as bit error rate (BER) back to "masks" that specify the maximum allowable envelope of each term of the form $\frac{a_i}{f^i}$, thus providing a framework for circuit design aimed at controlling the phase noise PSD.

Chapter 2

Mm-Wave Wireless Backhaul

2.1 Introduction

In this chapter, we consider a mesh network with highly directive mmWave links as a means of extending backhaul from wired gateways to the picocell access points. We formulate the wireless mesh backhaul design problem as a joint routing and resource allocation/scheduling optimization, with the goal of maximizing the access rate at the base stations, while accounting for the mutual interference between simultaneously active links. the highly directive nature of mmWave links leads to localized and sparse interference that we are able to exploit for efficient computation of the optimal solution. We illustrate our approach for two settings, urban and suburban, which exhibit interference patterns that are quite different.

We primarily develop a framework for *downlink* traffic routing and resource allocation, and then extend the formulation to perform joint uplink and downlink optimization. Additional flows between nodes in the network can be included in the optimization by introducing a new commodity for each flow. While we only consider fixed-rate traffic flows, extension to any concave objective of the rate vector is possible in the proposed

framework. The problem of resource allocation for optimizing backhaul throughput is initially formulated as a linear program, the dimensions of which grow exponentially with network size. Demonstrating the existence of a sparse solution to this linear program, we formulate an equivalent mixed-integer linear program that scales linearly with network size. The proposed formulation is able to solve relatively large networks with hundreds of nodes in a time scale of minutes.

2.1.1 Related work

Multihop wireless mesh networks have received extensive attention in the literature, typically with sensible heuristics for solving the associated joint routing and scheduling problems. Non-contention-based medium access protocols that rely on link scheduling and resource allocation have been studied for decades. Various works such as [18, 19] consider the link scheduling problem independently assuming routing is known beforehand, while others attempt to jointly optimize scheduling and routing. In modeling the interference behavior of links, the vast majority of these studies assume a binary effect, commonly known as the "protocol model" as described in [20]: Any two links either collide completely (mostly by contradicting the half duplex constraint or, in omnidirectional settings, falling within some interference radius) and have to be orthogonalized in resources, or have zero interference [21–30]. In [25, 26] a transmission is assumed to fail if the number of active transmitters within interfering range of its receiver is above a threshold, which is a modified version of the collision model.

Few papers take note of the residual interference between links in the network and allow for capacity-SINR trade-off in their allocation optimization; this can lead to suboptimal solutions and lost throughput as demonstrated in [31]. In [32], confining the analysis to the low SINR regime, link throughput is approximated as a linear function

of SINR. This approximation is not suitable for the high-speed point-to-point links employed for wireless backhauling.

Several papers consider the specific problem of multihop mesh backhauling with directional antennas, but limit the number of links on each node to a single steerable beam [4,21,25,26,28,33]. Other works such as [23,30] allow simultaneous multi-neighbor communication but designate a fixed discrete partitioning of resources, generally bandwidth, and limit each link to a single partition at any time. Coarse discretization of resources and single channel confinement of link transmissions both prevent achieving the true capacity of the network, and orthogonalizing same-node links severely limits the potential for spatial reuse, especially as such limitations are not present when employing highly directional antennas that reliably isolate many simultaneous same-node transmissions. In contrast, our work allows continuous partitioning and places no constraint on which part(s) of the resources each node may use other than what is imposed by the physical interference model, thereby realizing the true capacity of the network and facilitating the spatial reuse required for backhaul delivery.

Increasing cellular capacity through self-backhauled small cells has been the subject of many previous works. Some studies consider placement of relay nodes inside cells to improve signal quality at cell edges [24, 34, 35], yet the capacity boost obtained through simple radio frequency (RF) amplification and relaying is limited. Other studies consider addition of small-cell base stations inside a macrocell, each receiving wireless backhaul directly from the wired macro-BS, forming a backhaul network with star topology [4, 36, 37]. In [38], a multihop network is considered in the form of a row of nodes inside one street along with one wired node that acts as a gateway. In [24], the network comprises of one base station, relay nodes which do not generate data, and end nodes which do not relay data. These simplified structures are useful to provide insight into the capacity of wirelessly-backhauled picocells, but evaluating and optimizing a realistic network requires

increasing the scope to include multihop paths along different streets and networks with several gateways.

In contrast to the heuristics employed in prior literature on multihop mesh networks, we are able to compute the optimal solution by considering all possible link activation patterns. The idea of allocating resources to "cliques" of links has been around for decades [39], yet it has also been shown that finding the true optimal allocation and routing is an NP-hard problem the complexity of which scales exponentially with network size [40, 41]. Several studies tackle the intractable problem of interference-aware routing and resource allocation using various heuristics [21, 41–47]. These include either relaxing the interference model to a conflict graph and solving the resulting problem using edge coloring algorithms or column generation based methods that iteratively search for "good" patterns and solve the allocation problem over the discovered patterns in each iteration. These approaches often fail to obtain the exact optimal throughput due to their limitations in modeling or greedy procedure.

Finally, in [48], a mmWave multihop backhaul network was modeled as a uniform square grid of nodes to provide analytical insight into throughput capacity as a function of the size of the cluster supported by each gateway. An interesting observation was that backhaul capacity is not diminished as a result of residual interference, and, with careful scheduling of interfering links, the interference-free capacity of the network can be achieved. We demonstrate the importance of interference avoidance in scheduling and note that the proposed framework provides such a schedule for their specific topology.

Inspired by the approach used in [49–52] for optimizing downlink spectrum allocation in cellular networks, our starting point is the problem of resource allocation among all possible subsets of links. Then, following [53], we reformulate the convex combinatorial resource allocation problem into a scalable mixed integer optimization problem. The problem considered here is fundamentally different from that of [49–53] due to the mul-

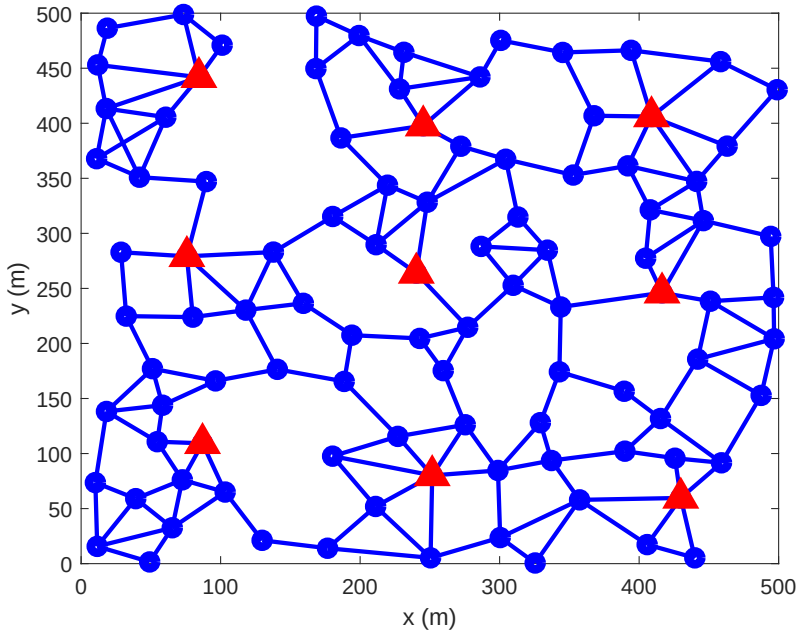
tihop nature of the network and the added problem of routing both uplink and downlink traffic.

2.2 System model

We envision an outdoor picocellular network with cell radii as small as tens of meters. The pico base stations are placed opportunistically on existing structures such as lampposts, building walls, and ceilings. Each base station, or "node", is connected to several nearby nodes (neighbors) through directional mm-wave links. Only a fraction of nodes are equipped with wired backhaul connection to the communication infrastructure; these nodes operate as gateways. The downlink and uplink data of each node is routed through a multihop path in the mesh to one or more of the gateway nodes. Examples of this structure are shown in Figure 2.1 where the gateways are marked by triangles and non-wired base stations by circles. Each line connecting two neighboring nodes represents two wireless links, one in each direction. Phased array antennas are utilized at either end of the link to maintain directional transmit and receive radiation patterns and reduce interference. We define the nominal link SNR and nominal link rate as the SNR and throughput of links in the absence of interference, and transmit power is controlled such that these values are the same for all links.

We denote by $\Gamma = \{1, ..., N\}$ the index set of all N non-gateway nodes in the network and by $\Lambda = \{1, ..., L\}$ the index set of all links in the network. As far as interference allows, any number of links connected to a node can be active simultaneously. However, because a transmitted signal is generally strong enough to saturate all co-located receivers, communication is half-duplex, namely, the links connected to one node can transmit simultaneously, or receive simultaneously, but at no time do some links transmit and some receive. As the opposing directions between two nodes are regarded as

(a)



(b)

Figure 2.1: Example of (a) a 48-node, 4-gateway portion of an imagined picocellular backhaul network in Manhattan, (b) a randomly generated suburban network of 100 nodes and 9 gateways in a 500 m×500 m area. Gateways are identified by triangles and non-gateway nodes by circles.

two separate links in the proposed optimization framework, the duplexing constraints can be easily incorporated by assuming infinite (or disabling) interference levels for such conflicting links.

In order to find the optimal allocation for interference management and routing, we apply a combinatorial approach where the available resources (e.g., time and/or frequency) are divided between different *subsets* of links with the objective of maximizing the minimum backhaul throughput delivered to all nodes in the network. A general scheduling framework can be defined as follows. The total time in each frame is divided into fragments of variable lengths and during each fragment (or slot) a certain combination of links are active. We denote each possible partitioning of links into on and off by an "activation pattern", which determines the amount of interference each link is subject to, and hence its data rate. The resource management problem is equivalent to allocating to every possible activation pattern an appropriate fraction of all resources (which may be none). For every possible activation pattern $A \subset \Lambda$, let $x_A \in [0, 1]$ denote the fraction of resources allocated to that pattern, which must satisfy the resource constraint,

$$\sum_{A \subset \Lambda} x_A = 1.$$

The problem of allocating resources to L links becomes that of allocating non-overlapping portions to the $2^L - 1$ non-empty subsets of links.

While the uplink or downlink rates on access links for a given *mobile* user can be bursty, and vary significantly across small timescales, the backhaul capacity demand at each base station, generated by aggregating its access link traffic, is smoother. Thus, our framework supports a constant demand at each base station over each optimization period, while allowing the flexibility of allocating resources differently across different base stations. Specifically, as we shall see, we may specify that base station i receive

backhaul capacity allocation $\alpha_i d$, where d is a quantity to be maximized by joint routing and resource allocation.

We consider two distinct deployment scenarios, modeling suburban and urban settings. In the suburban setting, antennas are placed on rooftops that are of relatively similar heights. As a result, the street geometry does not have a significant effect on the channel as links are assumed to be line-of-sight (LOS) without any blocking structures between antennas. The free-space propagation model is thus used to model the channel between different nodes and the only factors determining signal and interference strength are the radiation pattern of antennas and distances between nodes. In the high-rise urban setting, antennas are mounted on below-rooftop-level structures such as lamp-posts, traffic lights, and building walls. In this case a street-canyon model is considered wherein the channel is a combination of the LOS path and single-bounce reflections from the two canyon walls and ground. Transmit and receive radiation patterns and Fresnel (specular) reflection loss are accounted for in the model, as well as phase offset between propagation paths that is assumed to be uniformly random over the span of $[0, 2\pi]$ radians. Streets are assumed to be 25 m wide and the distance of each node from street wall and ground is chosen uniformly over $(4, 21)$ m and $(5, 8)$ m respectively. In both urban and suburban scenarios we assume nominal link SNR of 10 dB and each antenna is an array of 32 omnidirectional elements with half-wavelength spacing.

In simulations, we consider the topology of Figure 2.1a for the urban scenario, and randomly generated topologies similar to Figure 2.1b for the suburban scenario. These are generated by placing nodes uniformly at random in a square area with a minimum link distance of 10 m. For each non-gateway node, a random number is drawn between 3 and 5, then the node is connected to that number of its nearest neighbors with bidirectional links, and links that are longer than a threshold are dropped. If two links on a node are too close in the angular domain (within two beamwidths), the longer link is dropped.

The actual degree of a node may be larger than the number of neighbors it connects to, since additional links may be established by its neighbors. Each bidirectional link is actually two separate links individually indexed in the set Λ. The maximum link length threshold is tuned to obtain a well connected graph and is set high enough to ensure that no node is left disconnected from the network. Gateway nodes create connections with at most 6 closest neighbors whose distance is below the link establishment threshold. This is an appropriate design practice to increase the backhaul capacity of the network, as gateway links are traffic bottlenecks.

The number of gateways is chosen to be approximately 10% of the total number of nodes, and gateway nodes are defined by choosing the closest nodes to uniformly placed anchor points in the area. While this generating scheme obtains relatively realistic topologies, the occurrence of sub-par configurations that impose bottlenecks to service of some areas is possible, which can significantly decrease the max-min capacity of the network. In practice, node placements would be optimized to some extent to improve the capacity of the network and better performance than the simulated outcomes can be expected.

2.3 Combinatorial formulation of the allocation problem

In this section we formulate an optimization problem that maximizes the minimum backhaul throughput delivered to every node subject to interference and resource constraints.

When pattern $A \subset \Lambda$ is active, the throughput of link $l \in A$ is equal to,

$$\gamma_{l,A} = \log\left(1 + \frac{S_l}{\sigma_n^2 + \sum_{k \in A \setminus \{l\}} I_{k \to l}}\right) \tag{2.1}$$

where S_l is the signal power spectral density (PSD) of link l, $I_{k \to l}$ is the interference PSD from link k on link l, and σ_n^2 is the noise PSD. While $\gamma_{l,A}$ is left undefined if $l \notin A$ to avoid redundant parameters, it is equivalent (and natural) to think of it as equal to 0 if $l \notin A$. The total data rate of link l, i.e., data transferred in a unit of time, is therefore obtained by,

$$r_l = \sum_{A \subset \Lambda : l \in A} x_A \gamma_{l,A}.$$

2.3.1 Optimization of downlink only

In the downlink, each node is considered as a sink of its own cell's traffic. The gateway nodes are sources with no throughput constraint. Although this is a network of multiple "commodities", each intended for one non-gateway node and potentially delivered via a combination of multiple flows, it is equivalent to a network of a single commodity by using the following insight: We add a virtual source node representing the core (wired) network where all flows originate, with a link of unlimited capacity to each of the gateway nodes. Since this is the sole source node in the network, *any* data that flows into *any* node has invariably originated in this virtual node and is therefore downlink backhaul traffic. The reader is referred to [54] for a detailed discussion of network commodities. The downlink data rate delivered to node i, normalized to system bandwidth, is thus equal to

$$d_i = \sum_{l \in I_i} r_l - \sum_{k \in O_i} r_k \ , \qquad i \in \Gamma \tag{2.2}$$

where I_i is the set of links that flow into[1] node i and O_i is the set of links that flow out of it.

If allocations were in the frequency domain, the spectral efficiency parameters $\gamma_{l,A}$ would depend on the allocation variables x_A as the noise power may depend on the bandwidth. Time domain allocation thus simplifies the formulation.

The network utility is in general a function of the rate vector $[r_1, ..., r_L]^T$. Our formulation provides the flexibility to provide different backhaul throughput to different nodes by guaranteeing throughput $\alpha_i d$ to node i, with optimal allocation maximizing d. This allocation can be obtained by solving the following optimization problem:

$$\underset{[x_A],[r_l],[d_i],d}{\text{maximize}} \quad d \tag{2.3a}$$

$$\text{subject to} \quad \sum_{A\subset\Lambda} x_A = 1, \tag{2.3b}$$

$$r_l = \sum_{A\subset\Lambda: l\in A} x_A \gamma_{l,A}, \qquad l \in \Lambda \tag{2.3c}$$

$$d_i = \sum_{l\in I_i} r_l - \sum_{k\in O_i} r_k, \qquad i \in \Gamma \tag{2.3d}$$

$$d_i \geq \alpha_i d, \qquad i \in \Gamma \tag{2.3e}$$

$$x_A \geq 0, \qquad A \subset \Lambda \tag{2.3f}$$

where r_l is the data rate on link l. The weighting factor α_i can be used to provide larger backhaul throughput for high-traffic hotspots if needed. (In our simulations, we assume uniform traffic at all nodes and set $\alpha_i = 1, \forall i$.) Note that the node flow constraints only apply to non-gateway nodes. Relaxing the resource, rate, and throughput equalities, (2.3b), (2.3c), and (2.3d) to inequality $\leq$ is inconsequential.

Both the objective and constraints in the optimization are linear, therefore an optimal

[1]Not to be confused with $I_{k\to l}$ which denotes the interference of link k on link l.

allocation can be found by solving a linear program with around 2^L variables. Of course, if all of these possible activation patterns receive non-zero time allocation, the complexity of scheduling would grow exponentially with network size. However, there is always a sparse optimal solution to (2.3) in the following sense:

Theorem 1 *For a network of L links, N non-gateway nodes, and one or more gateways, there exists a solution* $[x_A]_{A\subset\Lambda}$ *to the downlink scheduling problem of (2.3) that is at most N-sparse, i.e.,* $x_A = 0$ *for all but (at most) N activation patterns A.*

Proof: Consider the N dimensional vector $\mathbf{d} = [d_1, ..., d_N]^T$ that is a feasible point for (2.3) achieved by the allocation $[x_A]$. For the entries of this vector, we have,

$$\begin{aligned} d_i &= \sum_{l\in I_i} r_l - \sum_{k\in O_i} r_k \\ &= \sum_{l\in I_i} \sum_{A\subset\Lambda: l\in A} x_A\gamma_{l,A} - \sum_{k\in I_i} \sum_{A\subset\Lambda: k\in A} x_A\gamma_{k,A} \\ &= \sum_{A\subset\Lambda} x_A \left(\sum_{l\in I_i\cap A} \gamma_{l,A} - \sum_{k\in O_i\cap A} \gamma_{k,A} \right). \end{aligned}$$

The maximum utility is determined by $\mathbf{d}$, which can be written as a convex combination of $2^L - 1$ vectors of dimension N as,

$$\begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix} = \sum_{A\subset\Lambda} x_A \begin{bmatrix} \sum_{l\in I_1\cap A} \gamma_{l,A} - \sum_{k\in O_1\cap A} \gamma_{k,A} \\ \sum_{l\in I_2\cap A} \gamma_{l,A} - \sum_{k\in O_2\cap A} \gamma_{k,A} \\ \vdots \\ \sum_{l\in I_N\cap A} \gamma_{l,A} - \sum_{k\in O_N\cap A} \gamma_{k,A} \end{bmatrix}.$$

Hence $\mathbf{d}$ resides in a polyhedron in $\mathbb{R}^N$ with up to $2^L - 1$ corner points. Noting the resource constraint, $\sum_{A\subset\Lambda} x_A = 1$, Caratheodory's theorem [8] states that $\mathbf{d}$ can be written as a convex combination of at most $N+1$ of these $2^L - 1$ points. Therefore there

exists an allocation $[x'_A]$ satisfying $\sum_{A\subset\Lambda} x'_A = 1$ with at most $N+1$ nonzero entries that yields the same node service vector as $[x_A]$. An optimal $\mathbf{d}$ cannot be an interior point of the polyhedron; otherwise one may strictly increase all of its elements to a boundary point, which increases the objective d. This implies that there exists an N-sparse optimal allocation, hence the proof of Theorem 1.

For the mmWave networks considered here, the interference matrix can be very sparse, resulting in many possible optimal allocations when we solve the linear program. By adding very small random fluctuations to the interference matrix, the solution can be made unique with high probability, and since a sparse solution is guaranteed to exist, we will find one using this trick.

One problem that arises when solving (2.3) is that the objective function does not penalize suboptimal routing as long as the delivered throughput is unchanged. As a result, a shorter path toward a node is not differentiated from a longer path with more hops, which can result in unnecessarily long paths and excessive latency and power consumption. This can be prevented by adding a linear term to the objective that implicitly penalizes delay. Assuming transfer of data on each link represents one unit of delay in the network, the sum of all link data rates can be taken as a linear proxy for delay and power consumption. This is done by changing the objective of (2.3a) to

$$d - \lambda \sum_{l\in\Lambda} r_l, \tag{2.4}$$

where the weighting factor λ is chosen to be small enough to ensure service rate, d, is always prioritized over the delay penalty term. It is shown in [7] that a sufficient condition to enforce this priority is

$$\lambda < \frac{1}{L\sum_i \alpha_i} \tag{2.5}$$

which, in the case of uniform service to all nodes ($\alpha_i \equiv 1$), simplifies to

$$\lambda < \frac{1}{LN}.$$

2.3.2 Joint uplink-downlink optimization

In this section, we extend the formulation of (2.3) to include both uplink and downlink. While the single commodity model no longer applies, we can add a second commodity representing uplink data that can originate at any non-gateway node but terminates at a virtual *sink* node connected to all gateway nodes through infinite-capacity links. For this commodity, all traffic will ultimately end up at the virtual sink (core network) and is therefore uplink data, meaning the difference between outgoing and incoming traffic at any node is the uplink throughput provided for that node. The downlink and uplink commodities must share network resources, therefore we formulate the joint optimization by defining two sets of link rate variables, $[r_l^{\mathrm{d}}]$ and $[r_l^{\mathrm{u}}]$, corresponding to downlink and uplink service rates, $[d_i]$ and $[u_i]$, that satisfy

$$\sum_{l \in I_i} r_l^{\mathrm{d}} - \sum_{k \in O_i} r_k^{\mathrm{d}} = d_i,$$
$$\sum_{l \in O_i} r_l^{\mathrm{u}} - \sum_{k \in I_i} r_k^{\mathrm{u}} = u_i.$$

The rate on each link will be the sum of the rate supporting downlink and uplink data, and (2.3c) will be modified to

$$r_l^{\mathrm{d}} + r_l^{\mathrm{u}} = \sum_{A \subset \Lambda : l \in A} x_A \gamma_{l,A}.$$

The optimal allocation is hence obtained from solving the following optimization problem, wherein the constants α_i and β_i determine the relative downlink and uplink traffic of different nodes.

$$\underset{[x_A],[r_l^{\mathrm{d}}],[r_l^{\mathrm{u}}],[d_i],[u_i],c}{\text{maximize}} \quad c \tag{2.6a}$$

subject to

$$\sum_{A\subset\Lambda} x_A \leq 1 \tag{2.6b}$$

$$r_l^{\mathrm{d}} + r_l^{\mathrm{u}} = \sum_{A\subset\Lambda: l\in A} x_A \gamma_{l,A}, \qquad l \in \Lambda \tag{2.6c}$$

$$\sum_{l\in I_i} r_l^{\mathrm{d}} - \sum_{k\in O_i} r_k^{\mathrm{d}} = d_i, \qquad i \in \Gamma \tag{2.6d}$$

$$\sum_{l\in O_i} r_l^{\mathrm{u}} - \sum_{k\in I_i} r_k^{\mathrm{u}} = u_i, \qquad i \in \Gamma \tag{2.6e}$$

$$d_i \geq \alpha_i c\,, \quad u_i \geq \beta_i c, \qquad i \in \Gamma \tag{2.6f}$$

$$r_l^d \geq 0\,, \qquad r_l^u \geq 0. \qquad l \in \Lambda \tag{2.6g}$$

For this formulation, Caratheodory's theorem cannot be applied to the node rate inequalities as easily as in Theorem 1. However, a similar argument can be made for the link rate vector $[r_1^{\mathrm{d}}+r_1^{\mathrm{u}}, ..., r_L^{\mathrm{d}}+r_L^{\mathrm{u}}]$ that is a convex combination of the 2^L-1 points (enumerated by A) in L dimensional space, $[\gamma_{1,A}, ..., \gamma_{L,A}]^T$. For any allocation $\mathbf{x} = [x_A]_{A\subset\Lambda}$ that is able to support rate vectors $\mathbf{r}^{\mathrm{d}}$ and $\mathbf{r}^{\mathrm{u}}$ and, equivalently, node downlink and uplink vectors $\mathbf{d}$ and $\mathbf{u}$, there exists an L-sparse allocation $\mathbf{x}'$ that provides the exact same link rates and uplink and downlink node service rates. Thus a solution with no more than L active patterns can be guaranteed to exist for any target rate vector pair.

While it is not of interest to the backhaul scenario, any flow between some source node and some sink node can similarly be incorporated in the formulation by introducing a

new commodity with its own rate variables, $\{r_l^{\mathrm{f}}\}$, and flow constraints for the sink, source, and other nodes as follows.

$$\begin{aligned}
&\sum_{l\in O_{\text{source}}} r_l^{\mathrm{f}} - \sum_{k\in I_{\text{source}}} r_k^{\mathrm{f}} = c_f, \\
&\sum_{l\in O_{\text{sink}}} r_l^{\mathrm{f}} - \sum_{k\in I_{\text{sink}}} r_k^{\mathrm{f}} = -c_f, \\
&\sum_{l\in O_i} r_l^{\mathrm{f}} - \sum_{k\in I_i} r_k^{\mathrm{f}} = 0, \qquad\qquad i \in \Gamma\backslash\{\text{source}, \text{sink}\}.
\end{aligned}$$

The rate variables would then be added to the LHS of (2.6c) to enforce sharing of network resources between flows.

The formulations presented in this section are suitable for relatively small networks. As L increases to even moderate sized values, the number of variables in the problem grows exponentially until the time or space (memory) complexity becomes unmanageable. In the next section, a scalable reformulation is developed.

2.4 A scalable reformulation

One characteristic of the network that can be leveraged to reduce the problem size is the localized nature of interference, which allows decoupling of constraints between distant areas of the network. We first define the neighborhood of link l as the set of links, Λ_l, that cause non-negligible interference on it, i.e., whose signal strength at the receiving end of l is above a threshold. Subsequently, the "local activation patterns" of link l are all possible subsets of its neighborhood, $B \subset \Lambda_l$. A local activation pattern B is in effect when all links in B are active *and* all links in $\Lambda_l\backslash B$ are inactive. By this definition, the spectral efficiency of a link only depends on the activation pattern of links in its neighborhood, or its *local* activation pattern. To maintain a pessimistic estimate of

throughput, the interference of all links outside the neighborhood are added to the noise and interference level when calculating throughput. In reality, many of the links outside the neighborhood will not be active, but we can ensure this worst-case assumption does not affect the result significantly by setting the threshold to be low enough. There is thus a trade-off between computational complexity and accuracy: Choosing a low threshold yields more exact results at the expense of increasing neighborhood size and enlarging the problem. In our simulations, we set the interference threshold to 3 dB below the noise level, assuming a nominal SNR of 10 dB. We find the disparity between the pessimistic and actual throughput to be less than 1% in all simulated cases.

Similar to (2.3), we define the *local* allocation variable x_l^B that is the resource allocated to local activation pattern $B \subset \Lambda_l$ of link l. When enumerating local activation patterns, the empty set is also counted since a nonempty global pattern may activate none of the links in one neighborhood. The data rate on link l is thus equal to

$$r_l = \sum_{B \subset \Lambda_l : l \in B} x_l^B \gamma_l^B, \quad l \in \Lambda$$

where γ_l^B is the spectral efficiency of link l under local activation pattern B, derived (pessimistically) as

$$\gamma_l^B = \log\left(1 + \frac{S_l}{n + \sum_{k \in B \setminus \{l\}} I_{k \to l} + \sum_{j \notin \Lambda_l} I_{j \to l}}\right).$$

Recall that Theorem 1 ensures the existence of an optimal solution to (2.3) that activates at most N patterns. We therefore consider a segmentation of the unit time frame to N slots, indexed by $M = \{1, ..., N\}$. The global slots are of variable lengths, denoted by $\{y_m\}_{m \in M}$, that satisfy the resource constrains, $\sum_{m \in M} y_m = 1$. Let P_m denote the *global* pattern activated in the m-th slot. Then $P_m \cap \Lambda_l$ is the corresponding *local*

pattern in the neighborhood of link l in the m-th slot. These local patterns may overlap and their union is P_m. We define the augmented local allocation variables $x_l^{B,m}$ as the resource allocated to local pattern B of link l in slot m. Evidently, for every $l \in \Lambda$, $m \in M$, and $B \subset \Lambda_l$,

$$x_l^{B,m} = \begin{cases} y_m & \text{if } B = P_m \cap \Lambda_l, \\ 0 & \text{otherwise.} \end{cases}$$

For the throughput calculations to hold, local patterns included in a global pattern must be consistent with each other. To enforce this constraint, we introduce a discrete activation parameter for each local pattern, denoted by $q_l^{B,m}$, which is a *binary* variable that takes the value of 1 when its corresponding local pattern is active in global pattern P_m and 0 otherwise. Activation is enforced by the inequality,

$$x_l^{B,m} \leq q_l^{B,m},$$

and consistency is enforced by limiting the sum of *incompatible* local patterns to 1, allowing at most one of them to be non-zero. The local patterns $B \subset \Lambda_l$ and $A \subset \Lambda_k$, corresponding to activation variables $q_l^{B,m}$ and $q_k^{A,m}$, are "compatible" if and only if,

$$A \cap \Lambda_l = B \cap \Lambda_k,$$

which means any active link in B that happens to be in the neighborhood of link k is also active in A, i.e., the two patterns do not impose contradictory activations on any links in the overlap of their neighborhoods. An example of consistent local patterns is shown in Figure 2.2 with neighborhoods depicted as sets and links as elements of these sets. Assuming gray squares are inactive links, local pattern A in the neighborhood of

link k is consistent with pattern B_1 in the neighborhood of link l but inconsistent with pattern B_2 in the same neighborhood, since link l_0 is inactive in A but active in B_2.

Using the binary activation parameters defined above, consistency of local allocations can be enforced by the inequality,

$$q_l^{B,m} + q_k^{A,m} \leq 1, \quad \forall A \cap \Lambda_l \neq B \cap \Lambda_k. \tag{2.7}$$

Thus the optimization problem of (2.3) can be reformulated as:

$$\underset{[x_l^{B,m}],[q_l^{B,m}],[y_m],[r_l],d}{\text{maximize}} \quad d \tag{2.8a}$$

subject to

$$r_l \leq \sum_{m \in M} \sum_{B \subset \Lambda_l : l \in B} x_l^{B,m} \gamma_l^B, \qquad l \in \Lambda \tag{2.8b}$$

$$\alpha_i d \leq \sum_{l \in I_i} r_l - \sum_{k \in O_i} r_k, \qquad i \in \Gamma \tag{2.8c}$$

$$\sum_{B \subset \Lambda_l} x_l^{B,m} \leq y_m, \qquad l \in \Lambda, m \in M \tag{2.8d}$$

$$\sum_{m \in M} y_m \leq 1, \tag{2.8e}$$

$$x_l^{B,m} \leq q_l^{B,m}, \qquad l \in \Lambda, B \subset \Lambda_l, m \in M \tag{2.8f}$$

$$q_l^{B,m} + \sum_{\substack{A \subset \Lambda_k \\ B \cap \Lambda_k \neq A \cap \Lambda_l}} q_k^{A,m} \leq 1, \tag{2.8g}$$

$$l, k \in \Lambda, B \subset \Lambda_l, m \in M$$

$$q_l^{B,m} \in \{0, 1\}, \qquad l \in \Lambda, B \subset \Lambda_l, m \in M \tag{2.8h}$$

$$x_l^{B,m} \geq 0. \qquad l \in \Lambda, B \subset \Lambda_l, m \in M \tag{2.8i}$$

In this formulation, r_l is the data rate of link l and (2.8c) is the set of flow constraints that

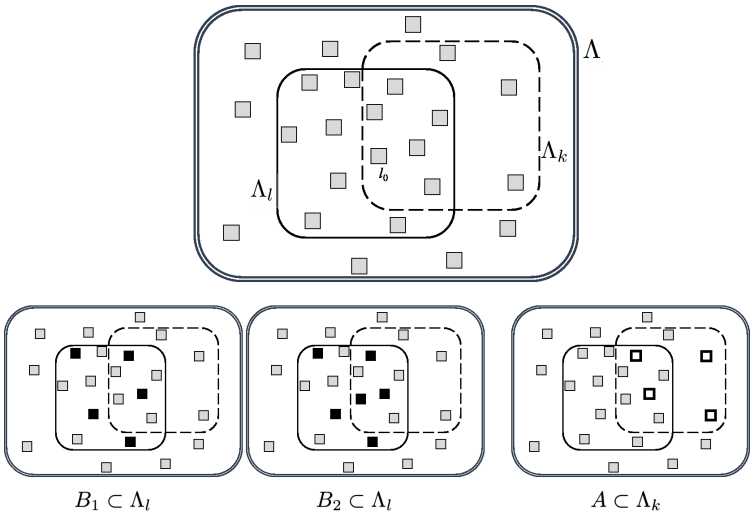


Figure 2.2: Examples of consistent and inconsistent local patterns: local pattern B_1 of link l (left) is compatible with local pattern A of link k (right), while local pattern B_2 (middle) is not. Using the binary activation variables, the conflict between B_2 and A is enforced by imposing the constraint $q_l^{B_2,m} + q_k^{A,m} \leq 1$. Gray squares correspond to inactive links, filled black squares correspond to links active in local patterns of link l, and white squares with black outlines correspond links active in local patterns of link k.

guarantees a minimum downlink throughput of $\alpha_i d$ to non-gateway node i. Equalities have been relaxed to inequalities in the rate, flow, and resource constraints, (2.8b), (2.8c), and (2.8e). Note that many of the consistency constraints of (2.7) have been bundled into a single inequality in (2.8g); this is possible because only one local pattern is active for each link in each slot, meaning these constraints can be compounded for local patterns of a single link.

The allocation that emerges from solving this problem is constructed as follows. The m-th global activation pattern is obtained by

$$P_m = \bigcup_{l \in \Lambda, B \subset \Lambda_l : q_l^{B,m} = 1} B,$$

and is alloted a time slot of length y_m normalized to the total frame. Using Theorem 1, it can be shown that any solution to (2.3) has an N-sparse equivalent that is further

equivalent, in terms of global patterns and allocations, to a solution of (2.8). We omit the proof and refer the reader to [50] for the proof technique.

The objective function in (2.8) can also be modified to incorporate delay penalization, by rewriting it as

$$d - \lambda \sum_{l \in \Lambda} r_l = d - \lambda \sum_{m \in M} \sum_{l \in \Lambda} \sum_{B \subset \Lambda_l : l \in B} x_l^{B,m} \gamma_l^B.$$

The second term is the sum of data rate on all links, and the weighting factor λ is chosen with the same threshold as derived for the combinatorial formulation in (2.5).

Both the objective and constraints of this formulation are linear, while the optimization variables are a mixture of continuous and discrete (binary) variables. Thus the exponentially growing linear program of (2.3) is reduced to a polynomially scaling mixed integer linear program. The number of variables in this formulation grows polynomially with network size due to the fact that as network size grows, *neighborhood* sizes remain the same. If the neighborhood size is no greater than μ, the number of local variables $x_l^{B,m}$ will be no more than $N \times L \times 2^\mu$. This brings the number of variables in (2.8) to a total of fewer than $N \times L \times 2^\mu + N + L + 1$ continuous and $N \times L \times 2^\mu$ integer values.

Similar to the combinatorial formulation, the above problem can also be modified to include both downlink and uplink by maintaining different link rate variables utilized for

the two directions of service. The resulting allocation problem is formulated below.

$$\underset{[x_l^{B,m}],[q_l^{B,m}],[y_m],[r_l^d],[r_l^u],c}{\text{maximize}} \quad c \tag{2.9a}$$

subject to

$$r_l^d + r_l^u = \sum_{m\in M} \sum_{B\subset\Lambda_l : l\in B} x_l^{B,m}\gamma_l^B, \qquad l \in \Lambda \tag{2.9b}$$

$$\sum_{l\in I_i} r_l^{\mathrm{d}} - \sum_{k\in O_i} r_k^{\mathrm{d}} \geq \alpha_i c, \qquad i \in \Gamma \tag{2.9c}$$

$$\sum_{l\in O_i} r_l^{\mathrm{u}} - \sum_{k\in I_i} r_k^{\mathrm{u}} \geq \beta_i c, \qquad i \in \Gamma \tag{2.9d}$$

$$\sum_{B\subset\Lambda_l} x_l^{B,m} \leq y_m, \qquad l \in \Lambda, m \in M \tag{2.9e}$$

$$\sum_{m\in M} y_m \leq 1, \tag{2.9f}$$

$$x_l^{B,m} \leq q_l^{B,m}, \qquad l \in \Lambda, B \subset \Lambda_l, m \in M \tag{2.9g}$$

$$q_l^{B,m} + \sum_{\substack{A\subset\Lambda_k \\ B\cap\Lambda_k \neq A\cap\Lambda_l}} q_k^{A,m} \leq 1, \tag{2.9h}$$

$$l, k \in \Lambda, B \subset \Lambda_l, m \in M$$

$$q_l^{B,m} \in \{0,1\}, \qquad l \in \Lambda, B \subset \Lambda_l, m \in M \tag{2.9i}$$

$$x_l^{B,m} \geq 0, \qquad l \in \Lambda, B \subset \Lambda_l, m \in M \tag{2.9j}$$

$$r_l^d \geq 0, \quad r_l^u \geq 0. \qquad l \in \Lambda \tag{2.9k}$$

Although the localized formulation is scalable in terms of problem size, unlike (2.4) (which is a linear program) this reformulation requires solving a mixed integer (binary) linear program which is inherently a combinatorial problem and NP-hard. While effective techniques have been devised for solving such problems, there are no guarantees for their computational efficiency. In the next section, some observations regarding scalability and

behavior of the two optimization frameworks of (2.3) and (2.8) are discussed.

2.5 Results and discussion

Solving the linear program formulated in (2.3) is relatively straightforward using standard techniques such as the simplex method. Standard solvers such as the CVX package in Matlab or Gurobi were used to solve this problem for small networks of up to 15 links. However, due to exponential growth of the problem size, solving a network with more than twenty links is practically impossible. In [7], this problem is sidestepped by clustering the network around gateways and solving each cluster independently. This provides a suboptimal solution wherein heuristics must be employed to determine cluster association.

The scalable formulation of (2.8), on the other hand, is a non-convex optimization problem due to the presence of integer variables. Approaches such as the branch-and-bound algorithm are generally effective methods for solving integer linear programs, but provide no guarantees for computational efficiency. For moderately large problems such as the 4-gateway network of Figure 2.1a, solving (2.8) using standard solvers (such as the CVX package in Matlab or the Gurobi solver) takes an impractical amount of time. However, we find that in practice, due to sparsity of the interference matrix, the number of global patterns that are activated in the solution can be much smaller than the upper bound guaranteed by Theorem 1, so that the number of global time slots can be set to a value $T \ll N$, indexed by the truncated set $M_T = \{1, ..., T\}$. Choosing an appropriate truncation level, T, involves a trade-off between computation time and throughput, as discussed in the next section.

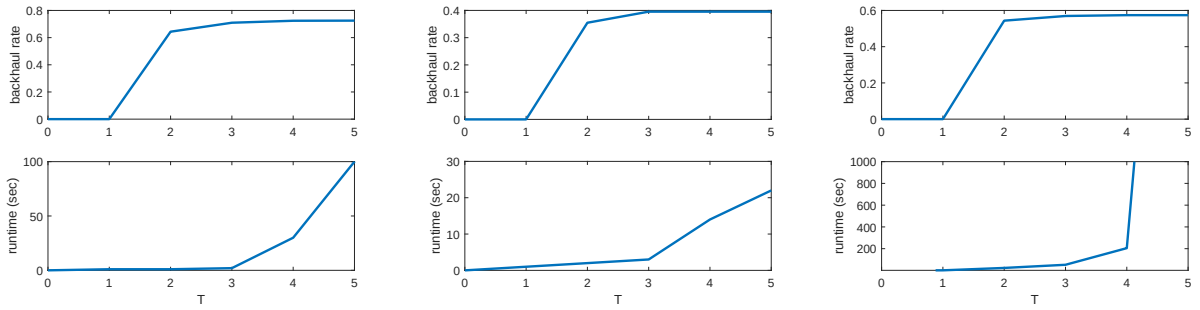


Figure 2.3: Effect of truncating M on obtained throughput and computation time in (a) a 4-gateway urban network with 48 nodes, (b) a 4-gateway suburban network with 50 nodes, and (c) a 9-gateway suburban network with 100 nodes. Nominal link rate is 3.46 (SNR=10dB). Depending on network topology, the minimum backhaul data rate delivered to every node is between 10% and 20% of nominal link rate.

2.5.1 Computational efficiency and scaling

Limiting the number of global activation patterns to a small number significantly reduces the computation time while achieving virtually all of the optimal throughput. It is also attractive in terms of implementation. Figure 2.3 depicts this trend, showing runtime and backhaul throughput obtained from solving (2.8) with different values for the number of global slots, T, in the urban network of Figure 2.1a as well as randomly generated suburban networks of 50 and 100 nodes similar to the structure depicted in Figure 2.1b. Note that the networks used in these simulations are randomly generated. Actual designed networks may have better characteristics in terms of node and gateway placement, as well as better choice of neighbor association, that prevent throughput bottlenecks and produce a more connected network. This would result in more uniform distribution of service among nodes and a higher max-min optimum throughput.

Figure 2.4 shows runtime as a function of network size for the original and scalable formulation (with the number of global slots, T, limited to 4). Networks of different sizes are generated using the suburban environment model. Formulation (2.3) blows up exponentially with network size, whereas the BLP of (2.8) can be used to optimize

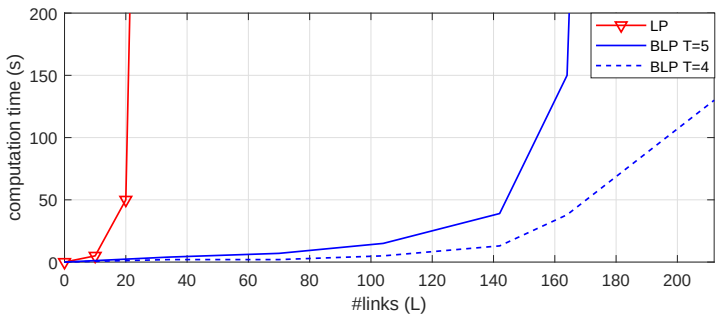


Figure 2.4: Computation time as a function of network size for combinatorial and truncated localized formulation.

networks of up to hundreds of nodes within a time-scale of minutes. In running the optimization, we found that by relaxing the rate equation of (2.8b) to inequality ($r_l \leq \sum_m \sum_B x_l^{B,m} \gamma_l^B$) computation time decreased considerably. The results presented here were derived with this relaxation.

2.5.2 Effect of residual interference

To quantify the effect of interference on backhaul capacity, we compare the throughput obtained in the two cases of (a) only modeling the half-duplex constraint (collocated TX-RX interference), and (b) including the full interference model described in Section 2.2, for the urban network of Figure 2.1a. We find that, similar to results reported in [48] and [7], the optimal throughput capacity is the same in both cases. In fact, the half-duplex nature of transmissions requires that any highly utilized (bottleneck) link at most be activated for a fraction of the time, so that the transmitted data can be *relayed* on the next link(s) on the multihop path in the remainder of the frame. This redundancy in link activation provides room for scheduling links such that no two interfering links are activated simultaneously. Thus backhaul capacity is not degraded from interference, but can only be obtained by careful scheduling of interfering links, and including interference in the model used for solving the allocation problem is crucial for obtaining such a

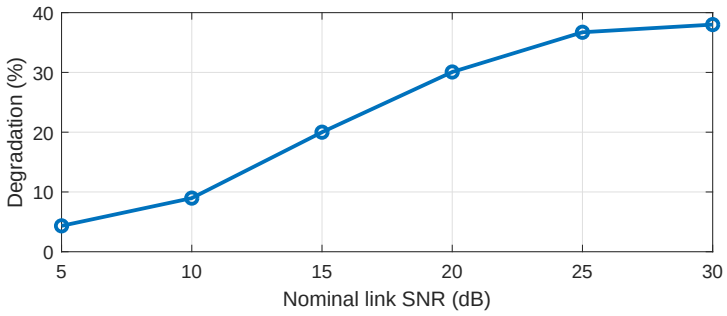


Figure 2.5: Degradation of throughput due to suboptimal interference-agnostic scheduling, as a function of nominal link SNR (typical example; numbers derived by evaluating (2.8) on the urban network of Figure 2.1a).

schedule. In fact, if an interference-agnostic allocation is deployed in the network of Figure 2.1a, the resulting max-min backhaul throughput is degraded by around 20% when evaluated in the presence of interference. This degradation becomes more severe as link SNR increases, as depicted in Figure 2.5.

2.5.3 Backhaul capacity in downlink and uplink

The backhaul throughput provided by the network differs depending on the density of gateway nodes, number of backhaul links connected to each gateway, network structure, and directionality of antennas. For a 10:1 ratio between non-gateway and gateway nodes, approximately 20% of the nominal link data rate can be delivered to nodes as downlink backhaul. Comparing with an existing network in which every base station is directly wired to the Internet, mm-wave wireless backhaul enables 10X shrinking of cell sizes by adding non-gateway base stations, resulting in significantly improved spatial reuse. The immense capacity of densely deployed picocellular access points predicted in [3] can thus be realized using wireless mesh backhauling, as long as the backhaul link data rates are high enough. For 10X increase in access point density, LTE cells with cell traffic in the order of hundreds of Mbps (or 1 Gbps with carrier aggregation) can be supported using

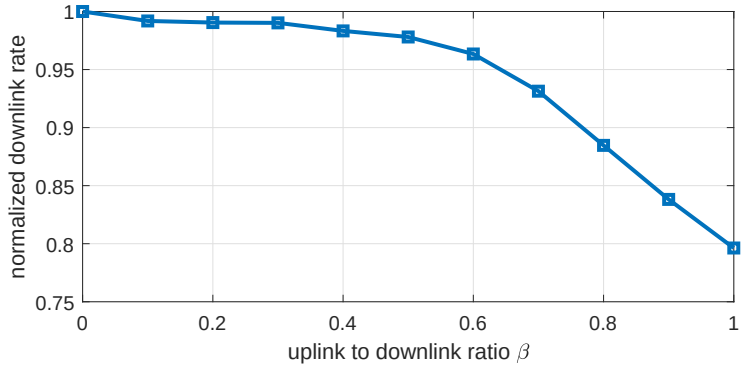


Figure 2.6: Reduction in downlink service rate as a function of uplink to downlink ratio. Results based on solving the urban network of Figure 2.1a with $T = 4$ global patterns.

wireless links with data rate of several Gbps, which is possible using the unlicensed 60 GHz band. On the other hand, high speed picocells that provide multi-Gbps mmWave-to-mobile access links may require tens of Gbps of backhaul throughput. This, in turn, would require backhaul links with raw data rates of the order of 100 Gbps, which could be realized using mm-wave or THz bands above 100 GHz.

To quantify the joint downlink and uplink throughputs, we fix the downlink ratios α_i to unity and solve for different values of uniformly distributed uplink ratios $\beta_i = \beta$. We observe that an uplink ratio of up to $\beta = 0.6$ can be supported with less than 5% degradation of downlink throughput. The trade-off between downlink and uplink rate is depicted in Figure 2.6 for the urban network of Figure 2.1a. Providing equal uplink and downlink capacity results in a throughput reduction of only 20% relative to downlink-only support. This is expected; the link capacity that is idle because of half-duplex relaying is effectively utilized by the uplink traffic that flows in the opposite direction, without much interference on the links carrying downlink traffic.

Figure 2.7: Example clusters formed by associating links to nearest gateway.

2.5.4 Cost of clustering

To demonstrate the benefit of optimizing a large multi-gateway network instead of independent optimization of single-gateway clusters, we manually divided the network of Figure 2.1a into four clusters as shown in Figure 2.7. Solving the entire 4-gateway network of Figure 2.1a results in per-node downlink rate of 21% nominal link rate, whereas solving for the clusters depicted in Figure 2.7 independently provides throughput of 17% to each node. Thus as network size grows, optimizing the network directly is preferable to clustering and provides higher throughput.

Chapter 3

Compressive User Tracking with Noncoherent Measurements

3.1 Introduction

In this chapter we propose methods for low-overhead tracking of the spatially sparse channel for large arrays with RF beamforming which is a critical challenge in realizing mobile mmWave links.

Conventional techniques for discovering spatial paths when constrained to RF beamforming include exhaustive and hierarchical scanning. The former has measurement complexity that scales linearly with array size, and is therefore not suitable for training massive arrays. In hierarchical scan, the number of measurements scales logarithmically with array size, but waiting for feedback from the receiver before each scan can be highly time-consuming and impractical upon implementation, and scaling with number of users is not optimal since each user may require a different beacon sequence. Compressive techniques, on the other hand, leverage the inherent sparsity of the mmWave channel to track users with a small number of measurements without suffering from the delay

of multiple feedback rounds. This approach can be implemented using RF beamforming with severely quantized phase control, which allows simplification of the RF front end.

Since current mmWave systems such as the 802.11ad standard are not designed to maintain phase coherence across packets, in this chapter we consider a *noncoherent* compressive estimation approach to develop an algorithm that can estimate a sparse spatial channel from RSS measurements of compressive beacons. For channels with a single dominant path, we derive and experimentally validate a noncoherent template matching scheme combined with Newton refinement for recovering the strongest path in the channel. We then propose a two stage algorithm for estimating the complete multipath channel. The first stage is phase retrieval, in which the phase of the RSS measurements is recovered up to a constant phase offset. The second stage is coherent compressive estimation on the the output of the phase retrieval stage. To adapt with simplified frontends with severely quantized beamforming weights, we choose one of the matrices, and infer the other via a pseudoinverse. We provide design guidelines for the (non-unique) virtual decomposition, and demonstrate its performance and scalability through simulations.

3.1.1 Related work

In [55], noncoherent tracking of multipath channels is attempted by designing beacon patterns that illuminate carefully chosen intervals of the angular space. Measurement of several such beacons can be used to identify the strongest paths in the channel. This method suffers from pattern imperfections caused by strong sidelobes that distort measurements. The sensing procedure is also disrupted by the possibility of destructive combination of paths that fall inside different bins in one beacon.

Our approach is inspired by recent work on compressive phase retrieval [56], which cascades phase retrieval with coherent compressed sensing to reconstruct a sparse signal

from noncoherent compressive projections. Our work differs from [56] in two key respects. First, rather than multiplying matrices known to be effective for phase retrieval and compressive sensing to obtain the measurement matrix, we provide a virtual decomposition that enables use of a measurement matrix that can be realized with coarse phase-only control. Second, we are interested in continuous-valued ("off-grid") parameter estimation rather than estimation of a signal that is sparse in a discrete basis, hence we replace the coherent compressive sensing stage by coherent compressive estimation.

Compressive recovery of sparse signals with incomplete measurements has been studied in several previous works. In [57,58], the problem of sparse signal recovery from 1-bit quantized projections is considered. In this case the sign of compressive projections is measured by the receiver and amplitude information is effectively lost. Such a problem can be formulated as a constrained $\ell 1$ norm minimization and solved efficiently. The loss of phase information, however, is more challenging. Several interesting approaches to sparse phase retrieval have been proposed in recent literature, including [59–61]. All of these methods are designed to recover a sparse vector and, when applied to path recovery on the continuum of spatial frequencies, suffer from grid mismatch error. This error can only be reduced by oversampling the spatial frequency continuum, which results in a larger "sparse" vector that requires more measurements to recover, effectively defeating the purpose of compressive estimation.

3.2 System model

We consider the link between a directional transmitter using a linear phased array antenna of size N (our approach applies directly to two-dimensional arrays, but we restrict attention to linear arrays for simplicity of exposition) and one or more receivers. The wireless channel of each link consists of a number of paths from the transmitter to the

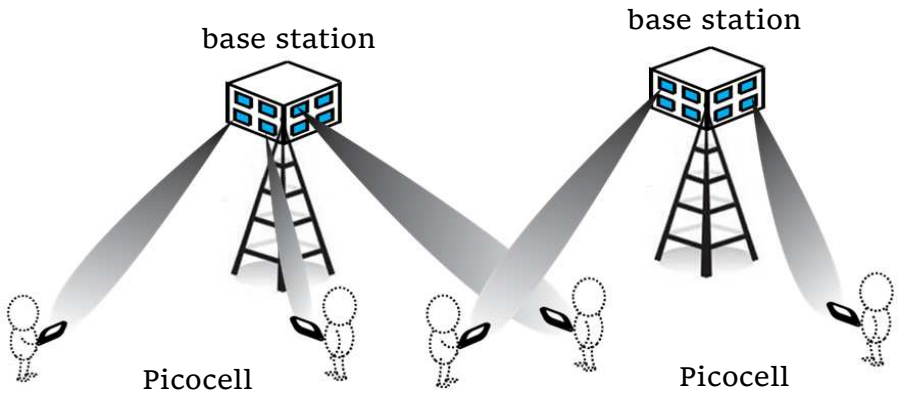


Figure 3.1: Base station to mobile communication using narrow pencil beams in the dense picocellular network.

receiver. The channel response is the sum of the responses of each of the paths on the array weighted by their respective complex amplitude. The array response of a path at angle of departure θ is equal to:

$$\mathbf{a}(\theta) = \left[e^{j1\frac{2\pi}{\lambda}d\sin\theta}, e^{j2\frac{2\pi}{\lambda}d\sin\theta}, ..., e^{jN\frac{2\pi}{\lambda}d\sin\theta}\right]^T$$

where d is the inter-element spacing of the array and λ is the wavelength.

Defining the *spatial frequency* $\omega = \frac{2\pi}{\lambda}d\sin\theta$, the array response of a path at angle θ can be described in terms of the corresponding spatial frequency ω by

$$\mathbf{a}(\omega) = [e^{j1\omega}, e^{j2\omega}, ..., e^{jN\omega}]^T$$

In the remainder of this chapter paths will be characterized by their spatial frequency instead of angle of departure. The net channel response on the N dimensional array is consequently equal to:

$$\mathbf{h} = \sum_{k=1}^{K} \alpha_k \mathbf{a}(\omega_k) \tag{3.1}$$

where α_k and ω_k are the complex amplitude and spatial frequency of the k'th path.

We assume that the transmitter broadcasts a series of M beacons, and each receiver

observes the strength of each beacon and provides an $M \times 1$ measurement vector as feedback to the transmitter at the end of the beaconing interval, which the transmitter employs to estimate the dominant spatial frequencies to the receiver. This procedure is depicted in Figure 3.2. Beacon b excites the array with the randomly generated weight vector $\mathbf{w}_b = [w_1^b, w_2^b, ..., w_N^b]^T$ which sprays the emitted power differently in different directions. The response of beacon b in the direction of spatial frequency ω is denoted by:

$$f_b(\omega) = \mathbf{w}_b{}^T \mathbf{a}(\omega). \tag{3.2}$$

The measurement made at the receiver will be a combination of the beacon response of all paths weighted by their corresponding complex path amplitude:

$$\begin{aligned} y_b = \mathbf{w}_b^T \mathbf{h} &= \sum_{k=1}^{K} \alpha_k \mathbf{w}_b{}^T \mathbf{a}(\omega_k) \\ &= \sum_{k=1}^{K} \alpha_k f_b(\omega_k) \end{aligned} \tag{3.3}$$

Therefore each angle of departure (AOD) will have its own M-dimensional beacon response "signature" and the weighted sum of the signatures of the paths in the channel arrives at the receiver. Based on the feedback from the receiver, the transmitter estimates the AOD of the strongest path(s) and beamforms in that direction for communication. Although the transmitter will generally beamform toward the strongest path in the channel, maintaining a library of multiple strong paths is useful for rapid recovery from sudden blockage of the strongest path, as well as for using alternate paths to manage inter-cell interference. Due to time-variant channel conditions and mobility of the receiver, channel estimation is repeated periodically to track the channel. This tracking takes place at a high enough rate to ensure the time between consecutive measurements is shorter than channel coherence time.

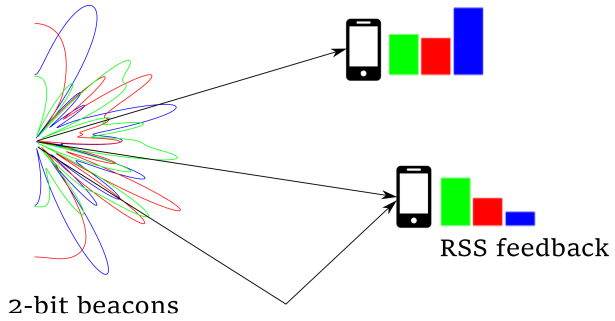


Figure 3.2: Radiation pattern of compressive beacons on a 16-element array excited by weights from distribution $U(\{\pm 1, \pm j\})$, and feedback of RSS measurements made by mobile users.

For simplicity, we assume omnidirectional reception, as in the early work on coherent compressive estimation [10]. Including the effect of receive arrays in noncoherent compressive estimation is an important topic for future work.

In our proposed system, compressive beacons are used for channel discovery with certain limitations on array weights. First, we assume no control over the amplitudes of the sensing matrix and confine the system to unitary weights. Second, we assume limited control over element phases to simplify the front end hardware; to this end, we allow 2-bit phase control for the sensing matrix so that array phases are selected from the set of $\{\pm 1, \pm j\}$. These limitations are crucial for minimizing hardware complexity as array sizes grow. As shown in [10], the degradation in beamforming performance caused by severe phase quantization is negligible as array size grows, meaning sacrificing phase granularity for array size can be an advantageous tradeoff.

Another important limitation is the loss of coherence from one beacon to the next. Due to oscillator design and dynamics, there is an unknown frequency offset between the local oscillators at the transmitter and receiver, that is constantly changing at an unknown rate. This random offset translates into a random offset in the phase of the channel measured by each beacon, corrupting the phase information of beacon measurements. For this reason, we assume RSS-only measurements are made at the receiver,

effectively changing the coherent measurement model of (3.3) to noncoherent magnitude measurements formulated by (3.4).

$$y_b = \left|\mathbf{w}_b^T \mathbf{h}\right| = \left|\sum_{k=1}^{K} \alpha_k f_b(\omega_k)\right| \tag{3.4}$$

The next sections present our proposed algorithm for compressive estimation of sparse channels under the nonlinear measurement model of (3.4) for single-path and multipath recovery.

3.3 Noncoherent compressive estimation of single-path channels

For a single-path channel with complex amplitude α_0 and spatial frequency ω_0, the noise-perturbed measurement model is described as

$$y_b = |\alpha_0 f_b(\omega_0) + n_b|$$

which in the high SNR regime can be approximated as

$$y_b \approx |\alpha_0 f_b(\omega_0)| + n_b$$

where n_b are independent Gaussian random noise variables. In this case, the likelihood function for measurement vector $\mathbf{y}$ is proportional to

$$\log p(\mathbf{y}|\alpha, \omega) = \sum_{b=1}^{M} \frac{|y_b - |\alpha f_b(\omega)||^2}{2\sigma_n^2} + C$$

where C is a constant indendent of α and ω. The maximum likelihood criterion for joint estimation of $|\alpha_0|$ and ω_0 is therefore

$$\begin{aligned}(|\hat{\alpha}_0|, \hat{\omega}_0) &= \underset{a>0,\omega}{\operatorname{argmax}} \sum_{b=1}^{M} |y_b - |a f_b(\omega)||^2 \\ &= \underset{a>0,\omega}{\operatorname{argmax}} \sum_{b=1}^{M} a^2 f_b^2(\omega) - 2a|f_b(\omega)|y_b.\end{aligned}$$

For a given ω, the value of a that maximizes the likelihood is $a^* = \frac{\sum_b |f_b y_b|}{\sum_b |f_b|^2}$. Denote the nominal RSS pattern for spatial frequency ω by $\mathbf{x}(\omega) = [|f_1(\omega)|, |f_2(\omega)|, ..., |f_M(\omega)|]^T$. Then we see that the optimal (maximum likelihood) estimate is given by the intuitively pleasing rule of choosing the spatial frequency whose nominal RSS pattern best matches the observed RSS pattern. Specifically, the cost function to be maximized is the squared normalized inner product of these nominal RSS patterns with the measured RSS pattern:

$$J(\omega) = \langle \frac{\mathbf{y}}{||\mathbf{y}||}, \frac{\mathbf{x}(\omega)}{||\mathbf{x}(\omega)||} \rangle^2,$$

where, for vectors $\mathbf{u}$, $\mathbf{v}$, $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T \mathbf{v}$ denotes inner product, and $||\mathbf{u}|| = \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}$ denotes the norm. Our estimate of the spatial frequency of the most dominant path is given by

$$\hat{\omega}_0 = \underset{\omega}{\operatorname{argmax}} \ J(\omega) \tag{3.5}$$

We solve this optimization problem by first evaluating the cost function on a discrete oversampled grid, and then refining around the maximizing grid point via Newton pursuit.

We now provide insight into the relative efficacy of the noncoherent and coherent frameworks, showing that, despite some performance degradation due to lack of coherence, the noncoherent framework can be expected to provide an accurate estimate for the dominant path. Ignoring noise, the RSS observations $\mathbf{y}$ are proportional to $\mathbf{x}(\omega_0)$, so

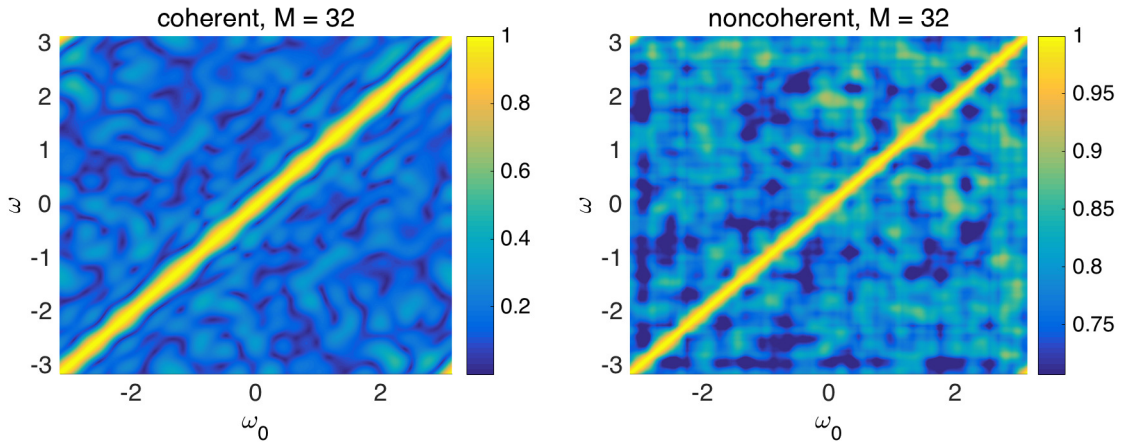


Figure 3.3: Correlation of beacon responses of different spatial frequencies using coherent and noncoherent measurements, with 32 beacons.

that the variations across ω are captured by the normalized correlations,

$$K(\omega, \omega_0) = \left| \langle \frac{\mathbf{x}(\omega_0)}{||\mathbf{x}(\omega_0)||}, \frac{\mathbf{x}(\omega)}{||\mathbf{x}(\omega)||} \rangle \right|.$$

Clearly, we will have a peak at $\omega = \omega_0$, but large local maxima at other values of ω would imply higher vulnerability to noise. We can also define analogous normalized correlations for (idealized) coherent measurements.

Figure 3.3 shows these normalized correlations $K(\omega, \omega_0)$ for both coherent and noncoherent measurements with 32 beacons. Clearly, coherent measurements provide better suppression of such undesired local maxima, but noncoherent measurements also provide enough discrimination between the desired $\omega = \omega_0$ line and undesirable local maxima.

3.4 Experimental validation of the single-path detection strategy

Using a 60GHz phased array testbed, provided by the Facebook Terragraph project [14], we perform initial validation of our noncoherent scheme in terms of accuracy and over-

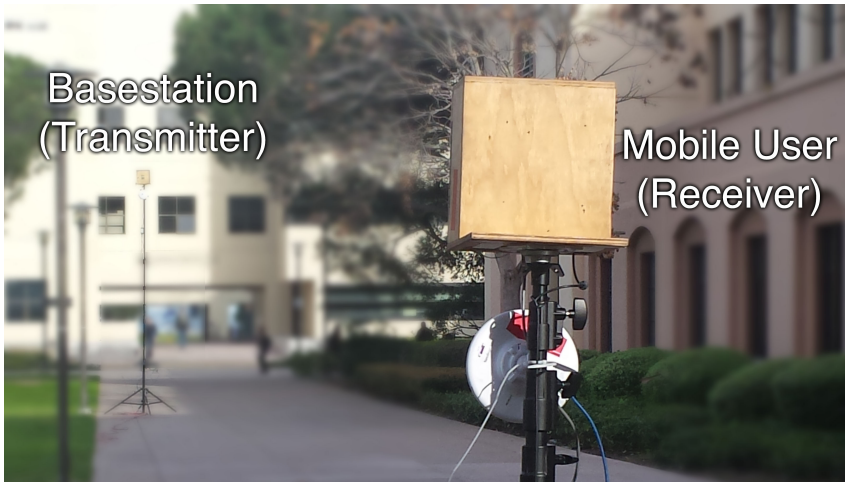


Figure 3.4: Our 60GHz phased array testbed emulating picocells.

head. Overall, our results show that our noncoherent design achieves accuracy similar to that of exhaustive beam scanning but at 50% smaller overhead. We also use simulations to show that this advantage increases sharply with the number of steerable antenna elements. Thus using commodity 802.11ad hardware, our design can lead to accurate user tracking at sub-second time scales even for very large arrays.

The 60GHz Phased Array Testbed. Our testbed consists of two 60GHz radios, each operating according to the 802.11ad single carrier mode [62]. To emulate a picocellular network setting, we place the transmitter (base station) at a height of 4 m, and mount the receiver on a mobile rack at 1.5 m emulating a user holding a smartphone (see Figure 3.4). Each antenna consists of an 8-by-16 rectangular array, with effectively 16 horizontally steerable elements (each consisting of 8 vertical elements with fixed phase relationships). The phase control uses a 4-bit quantization, allowing us to also perform experiments 2-bit phase quantization by using only 4 of the 16 available phase settings. Each radio is also paired with a WiFi radio for uplink feedback and system control. The base station broadcasts each beacon as a 802.11ad packet with zero payload. The receiver (user) operates in the omnidirectional mode to receive beacons.

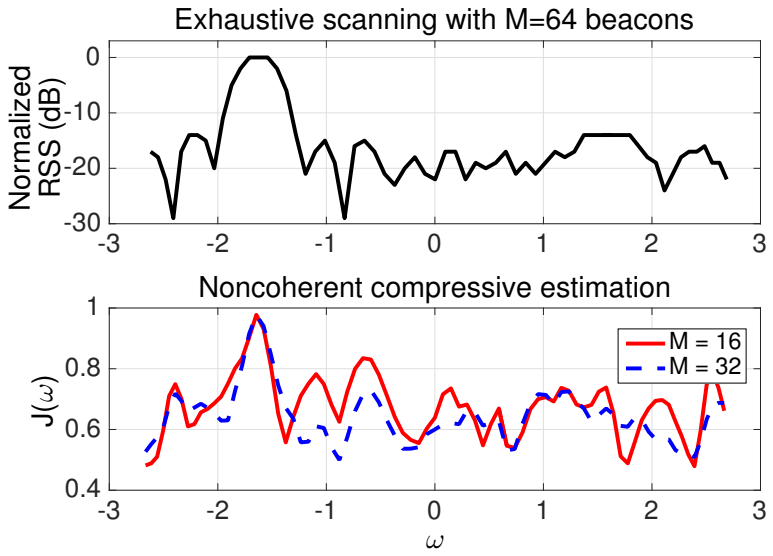
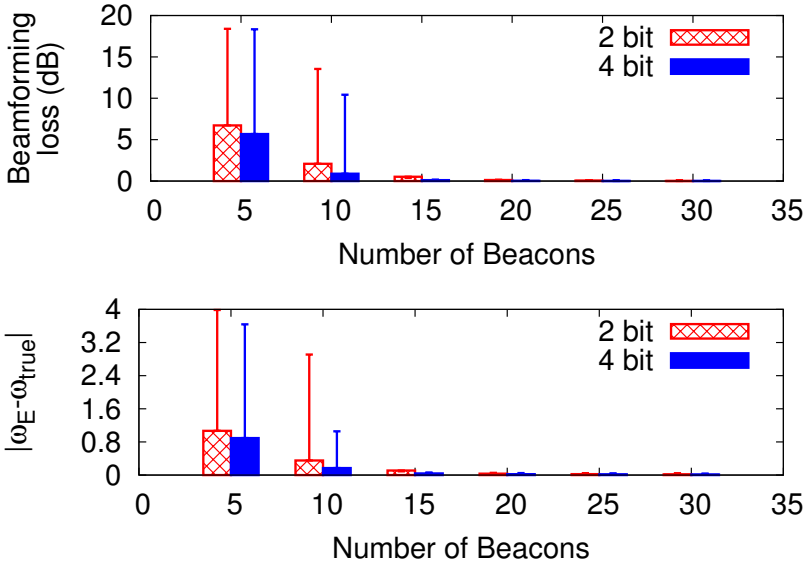


Figure 3.5: Exhaustive scan measurements vs. likelihood curve obtained by compressive measurements.

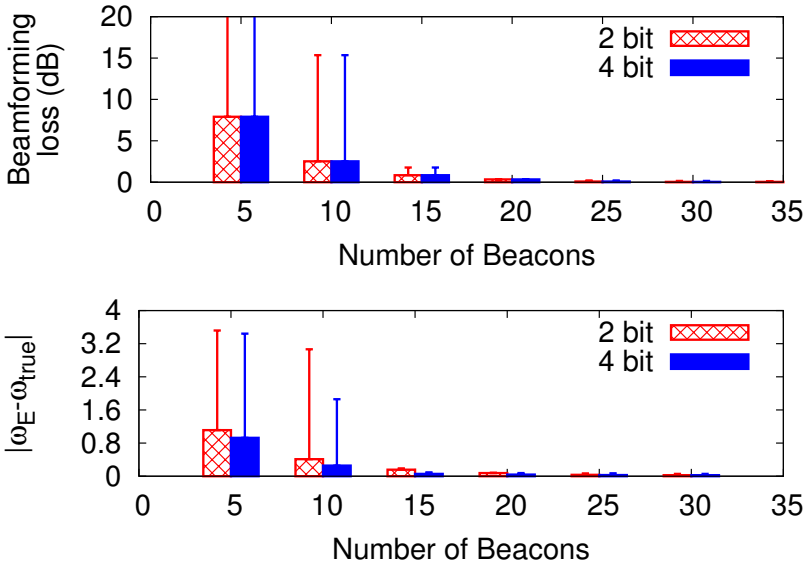
We performed experiments at various outdoor locations on our campus (open space, pathways near trees and buildings, etc). In most scenarios we could only find one dominant path. When placing antennas close to relatively smooth building surfaces we obtained two-path scenarios. We ran path estimation in three different ranges of 50, 100, and 200 meters, and observed similar results.

Path estimation accuracy We first evaluate whether our noncoherent compressive design can estimate the dominant path accurately. To obtain the ground truth path direction(s), we perform the exhaustive beam scanning of the azimuth angle with 64 beacons in the angular span of $(-45°, 45°)$ (the valid angular span for our hardware configuration), and obtain the corresponding RSS value per beacon. Under the same propagation environment, we run the compressive estimation using 5-35 beacons. From the user returned beacon RSS values, we generate the likelihood function $J(\omega)$ and identify the peak as the most dominant path.

Single Path Scenarios: Figure 3.5 shows an example result of exhaustive scanning (RSS vs. spatial frequency ω) with 64 beacons and the likelihood function $J(\omega)$ from non-



(a) Single path scenario



(b) Two path scenario

Figure 3.6: Mean & 95th percentile of beamforming loss and absolute spatial frequency error of noncoherent estimation vs. the number of beacons.

coherent compressive estimation with 16 and 32 beacons. We see that the peaks of $J(\omega)$ coincide closely with the peak of the RSS from exhaustive scan, demonstrating the accuracy of compressive estimation.

Next, we plot in Figure 3.6(a) the median and 95th percentile of the path estimation error (absolute spatial frequency error $|\omega_{\text{est}} - \omega_{\text{true}}|$), and the resulting beamforming loss (relative signal strength loss when the base station beamforms to the user in the estimated direction), across all measurements. Here we also examine the impact of phase control granularity, 4-bit (maximum available on the testbed) and 2-bit. We make two key observations. *First*, with just 15 beacons, our design achieves accuracy similar to that of exhaustive beam scanning with 32 or 64 beacons. *Second*, the estimation error and beamforming loss are similar for 2-bit and 4-bit phase control, confirming that our design can be implemented on low-cost hardware with coarser phase control.

2-Path Scenarios: In this case, the existence of a large flat surface (building wall or windows) provides modest reflection to the user. But the reflection path is always more than 8 dB weaker than the LoS path (due to reflection loss). Results in Figure 3.6(b) show that our noncoherent design can always detect the most dominant path at high accuracy.

Tracking overhead and latency. The overhead of path tracking includes two parts: the time taken by base stations to send beacons, and the time taken by users to feedback beacon RSS reports. The *beacon overhead* depends on the number of beacons and the beacon spacing, since each beacon is short (2μs). For our current 802.11ad hardware, the beacon spacing is bounded by the beam dwelling time (≤ 25 μs). Thus, the overhead for sending 16 beacons is 0.4 ms, when we adhere to the 802.11ad standard. The *feedback overhead* depends on the number of users and the wireless technique and protocol used for uplink. When using WiFi (802.11a/b), it takes on average 3.6 ms for one user and 30ms for five users (due to uplink contention). This delay can be significantly reduced

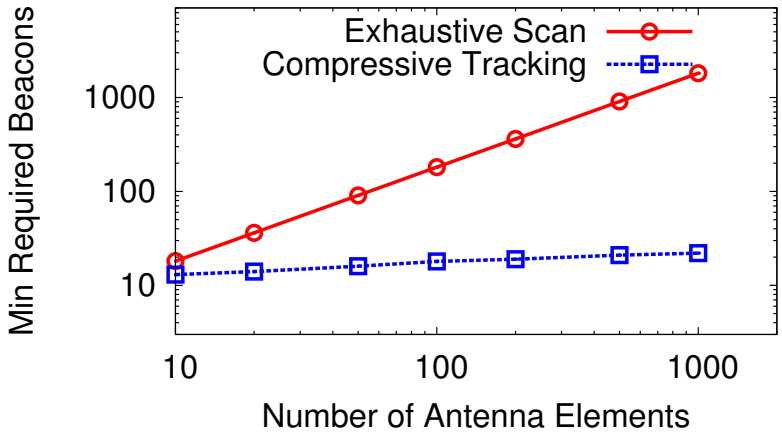


Figure 3.7: Beacon overhead vs. # of antenna elements

by improving the uplink design (which we leave to future work).

To evaluate the overhead when using large arrays, we use simulations to derive the minimum beacons required to ensure at most $3dB$ beamforming loss, as a function of the number of steerable antenna elements (N). Figure 3.7 shows that the beacon overhead for exhaustive scan scales linearly with N and that of our design scales nicely with $\log(N)$. For example, supporting arrays with 1000 steerable elements will require 22 beacons for our design and yet 1800 beacons for exhaustive scan. Using 802.11ad hardware, our design leads to 0.55 ms beacon overhead, or 0.55% overhead for a 100 ms frame, while exhaustive scan faces 45 ms overhead.

3.5 Noncoherent compressive estimation of multi-path channels

While several algorithms have been proposed for compressive estimation of sparse vectors, we focus on the Newtonized orthogonal matching pursuit method (NOMP) which is a robust, computationally efficient, and scalable estimation framework suited for sparse estimation on a continuum [63]. In order to maintain the benefits of NOMP and in-

spired by the approach put forth in [56], our method poses the estimation problem as a concatenation of phase retrieval and coherent compressive estimation via NOMP. The measurement matrix A is defined as the product of a phase retrieval matrix, A_{PR} of size $M \times M_{CS}$, and a compressive sensing matrix, A_{CS} of size $M_{CS} \times N$:

$$A = A_{PR}A_{CS}$$

The measurement vector is then denoted as

$$\mathbf{y} = |Ah|^2 = |A_{PR}A_{CS}\mathbf{h}|^2$$

where the channel $\mathbf{h}$ is defined in (3.1).

We introduce the auxiliary measurement vector $\mathbf{y}_{CS} = A_{CS}\mathbf{h}$. This vector, if retrieved, effectively provides (up to an unknown, and irrelevant, complex gain) coherent compressive measurements of the target channel $\mathbf{h}$ by sensing matrix A_{CS}. The algorithm is thus divided into two stages: the complex auxiliary measurement vector $\mathbf{y}_{CS}$ is first estimated from RSS observations $\mathbf{y} = |A_{PR}\mathbf{y}_{CS}|^2$ using a phase retrieval algorithm, and the resulting estimate is subsequently used for compressive estimation of $\mathbf{h}$ by sensing matrix A_{CS}. The two stages of the algorithm are summarized as follows.

Stage 1: Phase retrieval. While several different algorithms have been proposed for phase retrieval, each with different specifications and conditions, we consider the method proposed in [56] for our baseline analysis, since it has provable performance bounds. This method is shown to obtain the target vector up to a constant phase ambiguity with high probability under the condition that elements of the $M \times N$ sensing matrix A_{PR} are i.i.d. Gaussian and the number of RSS measurements, or M, scales as $n \log n$ for target vector size of n.

For evaluating the algorithm performance on simplified front ends with quantized beamforming phases, we use the more robust Wirtinger Flow algorithm [64] that utilizes gradient descent with careful initialization of the target vector to insure convergence to the global minimum. The reader is referred to these texts for a detailed description of the two methods.

Stage 2: Compressive estimation. The complex valued estimates obtained for $\mathbf{y}_{CS}$ are used in this step to solve for $\mathbf{h}$ the compressive estimation problem,

$$\mathbf{y}_{CS} = A_{CS}\mathbf{h}$$

in which $\mathbf{h}$ is a sparse combination of K continuous-valued spatial frequencies as expressed in (3.1). For self-contained exposition, we briefly review the NOMP algorithm that we use to sequentially extract the spatial frequency components. Iteration t consists of the following steps:

- Subtract the response of all paths extracted so far from the measurement vector to obtain the residual vector $\mathbf{y}_r$;

$$\mathbf{y}_r = \mathbf{y} - \sum_{k=1}^{t-1} \hat{\alpha}_k \mathbf{a}(\hat{\omega}_k)$$

- By matching the residual observation vector with the oversampled response dictionary, identify the strongest frequency in the remaining mixture and its correspond-

ing complex amplitude on the oversampled grid of spatial frequencies, Ω;

$$\hat{\omega}_t = \underset{\omega \in \Omega}{\operatorname{argmax}}\, G_r(\omega),$$
$$G_r(\omega) = |\mathbf{f}(\omega)^H \mathbf{y}_r|^2 / \|\mathbf{f}(\omega)\|^2$$
$$\hat{\alpha}_t = \mathbf{f}(\hat{\omega}_t)^H \mathbf{y}_r / \|\mathbf{f}(\hat{\omega}_t)\|$$

Where the notation $\mathbf{f}(\omega)$ represents the vector $[f_1(\omega), f_2(\omega), ..., f_M(\omega)]$ of responses to the M beacons as described in (3.2).

- Use Newton refinement to fine-tune the identified frequency with respect to the residual. Add the resulting frequency and its corresponding complex amplitude to the set of estimated paths;

- Use cyclic Newton refinement steps on all paths extracted so far to fine-tune frequencies and amplitudes with respect to the measurement vector $\mathbf{y}$.

The algorithm proceeds until the residual energy is lower than a threshold, indicating all dominant paths have been extracted. The reader is referred to [63] for a detailed discussion of implementation and performance guarantees of the algorithm.

To ensure successful channel recovery of a K-sparse channel, the sensing matrix A_{CS} must satisfy the restricted isometry property (RIP), i.e. provide sufficient separation between any two K-sparse vectors in the observation space. Random matrices have been shown to provide this condition with high probability under suitable circumstances. In particular, it has been shown previously in [65] that if the elements of the $M_{CS} \times N$ matrix A_{CS} are generated from an i.i.d. Gaussian distribution, then A_{CS} will satisfy the above property with high probability when M_{CS} (the number of compressive measurements) scales as $K \log N$. These predictions are useful guidelines in designing the algorithm pa-

rameters and creating the measurement matrices, as well as optimally allocating resources to the two stages of the process.

3.5.1 Generating the sensing matrices

The compressive sensing and phase retrieval matrices, A_{CS} and A_{PR}, must satisfy certain requirements to ensure the performance of each stage. At the same time, in order to be compatible with the hardware requirements described in Section 3.2, the combined matrix $A = A_{PR}A_{CS}$, must take values in the set $\{\pm 1, \pm j\}$. Generating the required matrices for hardware implementation and processing is therefore not trivial. In fact, identifying an $M \times M_{CS}$ and $M_{CS} \times N$ pair of matrices whose product, an $M \times N$ matrix, is in the quantized space is an overdetermined problem that cannot be solved exactly. We use the following procedure to minimize the distance between the product matrix and its quantized version, and treat this distance as a measurement error that can be compensated for by increasing the number of measurements and using stable phase retrieval and compressive sensing algorithms.

It is well known that i.i.d. complex Gaussian matrices have the necessary projection characteristics for phase retrieval and compressive sensing and are a good choice for A_{PR} and A_{CS}. In the first step, the elements of A are generated independently from a uniform distribution over the allowed values, i.e.,

$$A(i,j) \sim U(\{\pm 1, \pm j\}),$$

and A_{PR} and A_{CS} are then defined so that 1) their matrix product is as close as possible to A, and 2) they are i.i.d. complex Gaussian (exactly for one of the matrices, and approximately for the other).

After producing the sensing matrix A, we independently generate A_{CS} from an i.i.d.

complex Gaussian distribution:

$$A_{CS}(i,j) \sim CN(0, 2\sigma^2), \quad \sigma^2 = \frac{1}{2N}$$

The phase retrieval matrix is then defined as:

$$\begin{aligned} A_{PR} &= A{A_{CS}}^{+} \\ {A_{CS}}^{+} &= {A_{CS}}^{H} \left(A_{CS}{A_{CS}}^{H}\right)^{-1} \end{aligned} \tag{3.6}$$

where A_{CS}^{+} is the pseudoinverse of A_{CS}. We now argue that the elements of A_{PR} are also approximately i.i.d. with complex Gaussian distribution. Since the elements of A_{CS} are i.i.d. and zero mean, its rows are approximately orthonormal, which motivates the following approximation:

$$A_{CS}{A_{CS}}^{H} \approx I_{M_{CS}}.$$

Substituting in (3.6) we find that ${A_{CS}}^{+}$ is approximately equal to ${A_{CS}}^{H}$, so that the elements of the latter can be approximated as i.i.d. complex Gaussian. Element (i,j) of A_{PR} is therefore equal to:

$$A_{PR}(i,j) = \sum_{k=1}^{M_{CS}} A(i,k){A_{CS}}^{+}(k,j) \approx \sum_{k=1}^{M_{CS}} A(i,k)A_{CS}(j,k)$$

It is easy to see that the elements of A_{PR} are zero mean and uncorrelated. Since each is the sum of a moderately large number of zero mean, independent random variables, we can invoke the central limit theorem to argue that the elements of A_{PR} are jointly complex Gaussian, and therefore i.i.d.

The product $A = A_{PR} \times A_{CS}$ is then reevaluated and quantized to obtain the final sensing matrix.

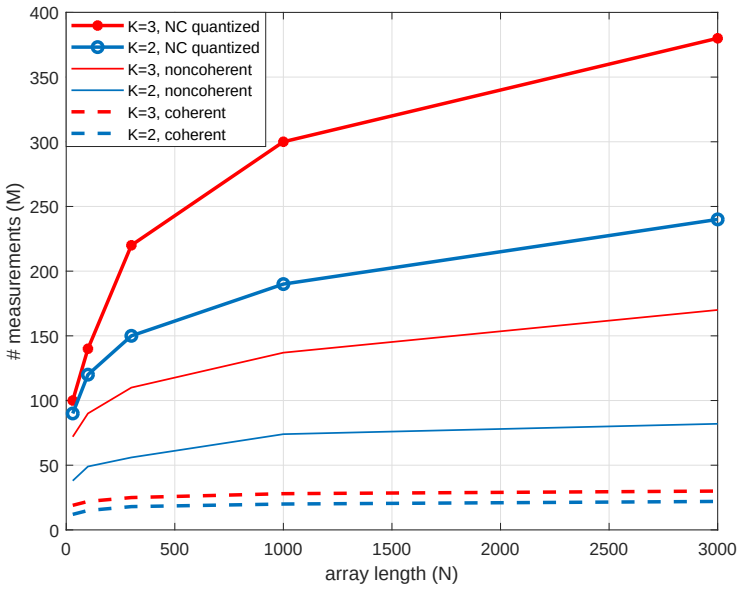
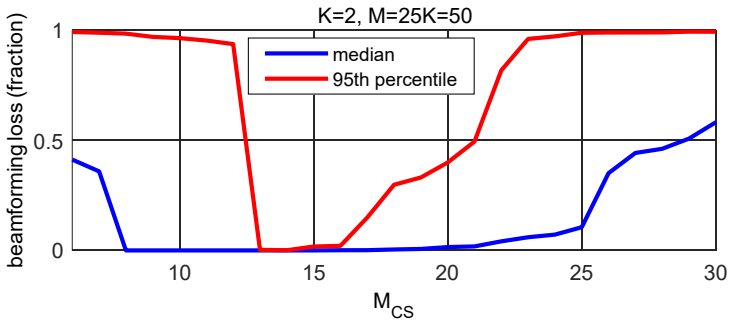
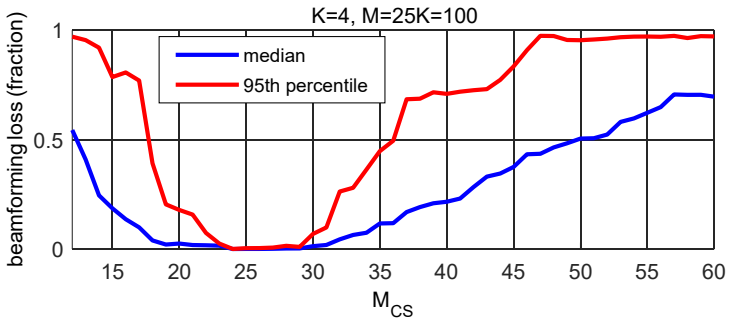


Figure 3.8: Required number of measurements M for 99% probability of accurate channel recovery (beamforming loss $\leq$ 1dB).



$K = 2$



$K = 4$

Figure 3.9: Beamforming error of estimated paths as a function of M_{CS} for a fixed observation size M. Array length $N = 1000$.

3.6 Numerical analysis of the multi-path channel estimation strategy

Both stages of the algorithm must succeed in order to accurately recover all strong paths in the channel. This requires appropriate choices for M and M_{CS}. The number of measurements required for coherent compressive sensing of an N-dimensional K-sparse vector is known to scale as $O(K \log N)$ [65]. The conditions for compressive estimation of a spatial channel with K significant components are more complex [66], but when the components are separated "well enough," the number of measurements scales with $O(K \log N)$ as well (Theorem 4 in [66]). On the other hand, successful phase retrieval also poses restrictions on the relation between M and M_{CS}. One fundamental lower bound for the number of observations required to identify M_{CS} complex values is $M \geq 2M_{CS}$ since each complex variable is defined by two real values. The best algorithms available until now perform this estimation using $M = O(M_{CS} \log M_{CS})$ real-valued observations [56].

Since M_{CS} scales as $K \log N$ and M scales as $M_{CS} \log M_{CS}$ (with the current state of the art), we conclude that the number of measurements required for the proposed algorithm scales with array size N and number of paths K as:

$$M = O(K \log N \log(K \log N)).$$

This is only slightly more costly than the $O(K \log N)$ measurement complexity of coherent compressive estimation. The proposed algorithm thus scales well with array size for sparse channels with a small number, K, of dominant components. Figure 3.8 depicts simulation results for the required number of measurements for accurate channel recovery as a function of array size for different values of K. These results are in agreement with the preceding theoretical predictions.

For a sufficiently large number of measurements M, the choice of M_{CS} must provide an optimum tradeoff between phase retrieval and compressive sensing accuracy. This tradeoff is demonstrated in Figure 3.9, where the overall performance of the algorithm is quantified for a fixed sensing configuration ($N = 1000, M = 25K, K = 2, 4$) as a function of M_{CS}. The best choice for M_{CS}, given by the lowest point of the curve, corresponds to the tradeoff between the accuracy of phase retrieval and compressive sensing that yields the best beamforming performance.

In Table 3.1 we provide a summary of performance characteristics for existing channel estimation algorithms alongside our proposed methods. In future work we aim to develop algorithms that are able to provide a library of viable paths (as the concatenated approach does) while also maintaining robustness to hardware constraints and weight quantization.

Table 3.1: Overview of channel sensing algorithms with RF beamforming.

	scaling	feedback overhead	RSS only	backup paths	quantized weights
Exhaustive scan.	N	low	yes	yes	OK*
Hierarchical scan.	$K \log N$	high	yes	yes	OK*
Compressive (NOMP)	$K \log N$	low	no	yes	robust
Noncoh. temp. match.	$K \log N$	low	yes	no	robust
PR/NOMP	$\sim K \log N$	low	yes	yes	vulnerable
???	$K \log N$	low	yes	yes	robust

*For large arrays, weight quantization has negligible effect on beamforming capability. However, the progression of *hierarchical* scanning is sensitive to beam patterns and can be degraded by weight quantization.

Chapter 4

Multi-Gbps Massive MIMO in the Presence of Phase Noise

4.1 Introduction

Most prior research in mmWave systems assumes RF beamforming, which employs one RF chain for the entire array, or hybrid beamforming, which employs a number of RF chains which is much smaller than the number of antenna elements in the array. However, advances in silicon implementations of mmWave hardware imply that, at least for a moderate number of antenna elements, it is possible to build low-cost RFICs with one RF chain for each antenna, opening up the possibility of all-digital beamforming for multiuser MIMO. In this chapter, we investigate modular architectures using such RFICs as "tiles," in a regime where the number of antennas per tile is fixed, but the size of the overall antenna array is scaled up by increasing the number of tiles. We consider a mmWave massive MIMO uplink, shown in Figure 4.1, in which the number of simultaneous users scales with the antenna array size, and our goal is to understand whether phase noise is a bottleneck for scaling the multiuser system.

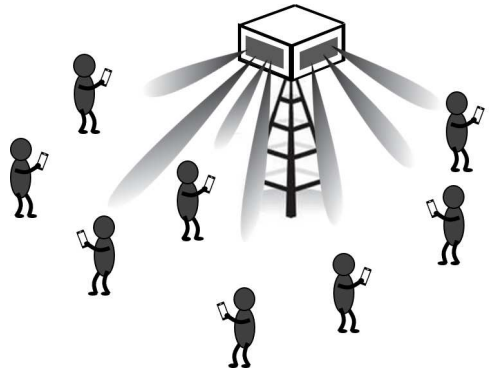


Figure 4.1: Mmwave multiuser massive MIMO.

While additive noise (and multiuser interference) can be averaged out (or suppressed) using degrees of freedom, the multiplicative nature of phase noise leads to distortion that scales with signal power, creating performance floors that can only be alleviated by reducing oscillator noise. Since realizing low-noise oscillators is challenging at higher carrier frequencies, we are interested in how much we can relax phase noise specifications while attaining a target system-level performance. To this end, we develop a framework for analyzing the impact of phase noise in massive MIMO, modeling its propagation in the proposed tiled architecture. Each tile is controlled by a separate RF chip that performs down conversion as well as analog-to-digital conversion, alleviating the need to transport analog RF signals across the entire frontend. To synchronize the tiles to emulate a single large array, a low-frequency reference is distributed to the tiles and multiplied up to the carrier frequency via a PLL on each tile. Thus, the sources of phase noise in this architecture are from the common low-frequency reference, and from the VCOs driving the PLLs in the tiles.

Concept system: While the analytical framework presented here is general, our numerical evaluations are based on a concept system operating at a carrier frequency of 140 GHz, with common low-frequency reference at 10 GHz. We consider single carrier QPSK modulation at 5 Gbaud symbol rate, corresponding to an uncoded bit rate of 10

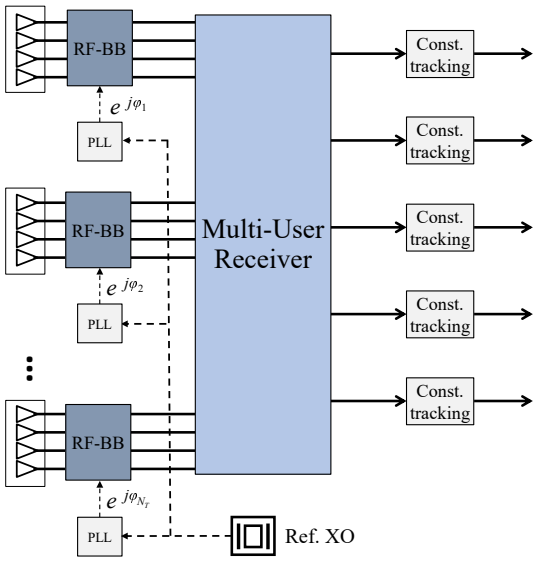


Figure 4.2: Architecture of the tiled multiuser massive MIMO receiver.

Gbps per user, and $N = 256$ base station antennas per sector. The load factor β, defined as the ratio of number of simultaneous users to the number of antennas, varies from $\beta = \frac{1}{16}$ to $\beta = \frac{1}{2}$ (with a nominal value $\beta = \frac{1}{4}$), which corresponds to sector-level uncoded throughputs ranging from 160 Gbps to 1.28 Tbps. These aggressive specifications may well be beyond what is required in deployed systems, but they allow us to explore the limits of system performance with reasonable hardware requirements. For the load factors of interest, spatial matched filtering does not yield acceptable performance, hence we consider LMMSE reception, arguably the simplest multiuser detection strategy we could adopt in exploring these limits.

We consider the architecture depicted in Figure 4.2, and model the propagation of phase noise through each of the blocks depicted therein, providing simple yet accurate estimates of its impact on system performance. The resulting analytical framework is used to provide design guidelines that greatly simplify the complicated task of joint hardware/system development. This includes determining maximum phase noise PSD

masks for oscillators that guarantee the desired performance level for a given system configuration. A key conclusion from the scaling laws we derive is that, for a given oscillator quality, phase noise is *less* harmful for massive MIMO than for a single-input single-output (SISO) system, and does not represent a bottleneck for our ambitious system goals.

4.1.1 Related Work

The effect of various hardware impairments (such as amplifier nonlinearity, low precision ADC and DAC, I-Q imbalance, and phase noise) on massive MIMO systems has been the subject of many studies, including [67–71] to name a few. Among these effects, phase noise is particularly challenging due to its multiplicative nature, especially for large communication bandwidth. Significant efforts have been made by the hardware community to extract accurate models for phase noise from the physics of oscillator circuitry. Various works have utilized the framework established in [72] to develop descriptions for the phase noise generated in different configurations [73–78], typically by describing the PSD of the phase noise process. Modeling the overall impact of phase noise on the communication link requires a system-level analysis that incorporates such models into the signal reception and decoding process. Prior studies in this direction, however, have often employed oversimplified models such as the pessimistic Wiener process model considered in [79] or the white noise model assumed in [80]. Many studies neglect the possibility of leveraging time domain correlations in a phase noise process for suppressing phase drift, unnecessarily tying system performance to channel tracking overhead by relying on frequent CSI updates for drift suppression [81], but tracking phase drift over time is crucial in realizing the true capacity of a practical system [82]. Furthermore, most studies that consider multiple antenna systems assume either a common clock with fully correlated

phase noise across the array (synchronous clock distribution), or a free-running oscillator at each element untethered to a common reference (asynchronous distribution), producing uncorrelated phase drifts at different antennas [80, 82–85]. Synchronous clocking is impractical at high carrier frequencies while the asynchronous configuration suffers from beamforming degradation as a result of rapid "channel aging".

In terms of modeling, the closest approach to ours in the literature is that of [86], which investigates the effect of phase noise on OFDM multi-user beamforming arrays. They assume an independent oscillator at each array element locked to a common lower frequency reference via a PLL multiplier, and model the filtering effects of the PLL on common and independent phase noise. In their analysis, the authors rely on a subset of subcarriers acting as pilots for phase noise tracking, and derive SINR predictions for single user and multiuser beamforming that predict similar scaling laws as our analysis. However, as shown in this and other studies, phase noise poses an inherent challenge for OFDM systems: the inter-carrier interference caused by phase noise increases with the number of subcarriers posing a fundamental limit on the bandwidth of an OFDM system impacted by phase noise [86–95]. OFDM also has other drawbacks for mmWave systems: the linearity required to handle high peak-to-average ratios is difficult to realize at reasonable power efficiency at high frequencies, and the precision required in analog-to-digital conversion is a challenge at large bandwidths. We therefore focus on a single carrier system in this chapter.

4.2 System model

We consider uplink multiuser MIMO in which the base station, equipped with an N-element digitally steered array, simultaneously receives signals sent by K users. The *load factor* is defined as the ratio of users to array size, and denoted by $\beta = K/N$.

We consider line-of-sight (LoS) channels between users and the base station. We focus on the impact of phase noise on multiuser demodulation rather than on channel estimation. Thus, we consider durations over which the spatial channels are well modeled as constant, and assume that the spatial channels at the beginning of such durations are known to the receiver. The vector $\mathbf{h}_k$ represents the channel of user k, with complex amplitude α_k and angle of arrival θ_k. For ease of notation we define the *spatial frequency* corresponding to θ_k as $\omega_k = (2\pi d/\lambda)\sin\theta_k$, where d is the array inter-element spacing and λ is the carrier wavelength. User k's channel can thus be represented by the sinusoid,

$$\mathbf{h}_k = g_k \left[\, 1, e^{j\omega_k}, e^{j2\omega_k}, \ldots, e^{j(N-1)\omega_k} \,\right]^T.$$

For simplicity, we assume that the K spatial frequencies are distributed uniformly over $(-\pi, \pi)$ (rather than uniformly over angles of arrival). In order to limit the variation in performance across users, we assume that users that are too close in spatial frequency are orthogonalized in time or frequency domain. and maintain a minimum pairwise distance of $2\pi/N$ in spatial frequency between users. While our analytical framework easily accommodates variations in user power, we assume perfect power control for simplicity of exposition.

For the concept system described in Section 4.1 our nominal configuration is $K = 64$, or $\beta = 1/4$. We consider single-carrier digital modulation with Gray coded QPSK, and use uncoded BER of 10^{-3} as our performance target, since low frame error rates can be achieved using high-rate error correction codes at this BER. We assume here that the bandwidth B (5 GHz for our concept system) equals the symbol rate $1/T_{\text{symb}}$, but the analysis easily extends to accommodate excess bandwidth. Nominal beamformed SNR (not including phase noise) is 14 dB. This provides a margin of 4 dB compared to the 10 dB required SNR for 10^{-3} BER for a SISO link without multiuser interference or

phase noise. LMMSE reception is used to separate user data streams, and constellation tracking with window size of 10 is performed at the output of each channel to offset slow-time-scale oscillator phase drift.

While we use uncoded BER as our metric here, our SINR-based analytical framework easily extends to alternative metrics such as spectral efficiency.

The modular architecture. Figure 4.2 depicts the modular structure of an N-element array containing N_T tiles, each with N_0 elements, so that $N = N_T N_0$. Our nominal configuration is $N_T = 16$ tiles and $N_0 = 16$ elements per tile. We define the "underloaded" case where the number of users is no larger than the tile size, i.e., $K \leq N_0$ or equivalently $\beta N_T \leq 1$. Since our goal is to scale up the system $(K, N \to \infty)$ while keeping tile size N_0 and load factor β constant, the underloaded regime will not be the operating condition of a large system and is of limited interest for scaling.

A 10 GHz reference is distributed to tiles and frequency multiplied on-tile using a PLL-controlled VCO to produce the 140 GHz carrier. Since the phase noises at different VCOs are independent, the carriers at different tiles contain independent phase noise components. The system block diagram for this process is depicted in Figure 4.3. In our running example, the multiplication factor is $N_f = 14$, producing a 140 GHz carrier from the 10 GHz reference clock. The PLL is type-2 with loop resonance frequency of $\omega_n = 1$ MHz and damping factor of $\xi = 0.707$, achieved by setting

$$k_V = N_f \omega_n^2, \qquad H_{\mathrm{LP}}(s) = \frac{1 - \frac{2\xi}{\omega_n} s}{s}.$$

Receiver modeling. In the absence of phase noise, the complex baseband signal received on the N-dimensional array is described by

$$\mathbf{x} = \mathbf{H}\mathbf{s} + \boldsymbol{\nu}, \tag{4.1}$$

where $\mathbf{H} = [\mathbf{h}_1 ... \mathbf{h}_K]$ is the $N \times K$ channel matrix, $\mathbf{s}$ is a K-dimensional vector containing the symbols transmitted by users, and $\boldsymbol{\nu} \sim \mathcal{CN}(\mathbf{0}, \sigma_\nu^2 \mathbf{I}_N)$ is the additive receiver noise vector. A linear receiver, $\mathbf{\Gamma}_{K \times N}$, is used to estimate the transmitted symbols as

$$\mathbf{y} = \mathbf{\Gamma}\,\mathbf{x}. \tag{4.2}$$

In the presence of phase noise, the received signals on separate subarrays are distorted by different phase noise terms during down conversion. Denoting by $\mathbf{H}_i$ the $N_0 \times K$ channel matrix of the i'th subarray, we have

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_1 \\ \mathbf{H}_2 \\ \vdots \\ \mathbf{H}_{N_T} \end{bmatrix}$$

and can model phase distorted reception by

$$\mathbf{y} = \mathbf{\Gamma}\left(\begin{bmatrix} \mathbf{H}_1 e^{j\phi_1} \\ \mathbf{H}_2 e^{j\phi_2} \\ \vdots \\ \mathbf{H}_{N_T} e^{j\phi_{N_T}} \end{bmatrix} \mathbf{s} + \boldsymbol{\nu} \right). \tag{4.3}$$

Since the additive noise, $\boldsymbol{\nu}$, is a vector of i.i.d. symmetric complex Gaussians, its distribution is not affected by phase noise, therefore we use the same symbol to represent the phase distorted version.

Assuming phase noise terms are small (which we ensure by design for the regimes of

interest), we use the approximation $e^{j\epsilon} \approx 1 + j\epsilon$ to write (4.3) as

$$\mathbf{y} \approx \mathbf{\Gamma}(\mathbf{H}\mathbf{s} + \boldsymbol{\nu} + \begin{bmatrix} \mathbf{H}_1 j\phi_1 \\ \mathbf{H}_2 j\phi_2 \\ \vdots \\ \mathbf{H}_{N_T} j\phi_{N_T} \end{bmatrix} \mathbf{s}). \tag{4.4}$$

Thus we see that phase noise introduces an additional distortion term to the classical multiuser reception model described by (4.1) and (4.2), which we model in detail in later sections.

4.3 Phase noise modeling

A perfectly noiseless carrier has constant complex baseband amplitude, $C(t) = A$. In a noisy setting, the complex envelope of an oscillator output becomes $C(t) = A + n(t) = A + n_c(t) + jn_s(t)$, where $n(t)$ is a complex Gaussian random process and n_c and n_s are its real and imaginary parts. For $A \gg |n(t)|$, we obtain our standard phase noise model

$$C(t) = Ae^{j\phi(t)}$$

where $\phi(t) = \frac{n_s(t)}{A}$ is a Gaussian random process with power expressed in dBc, or dB relative to carrier power.

As is conventional in the hardware literature, we denote the PSD of ϕ as $L(f)$. The dynamics of active components in oscillators produce *colored* phase noise in the output sinusoid [72–74]. This phase noise is modeled as a combination of white noise and lowpass components with PSD proportional to $1/f$, $1/f^2$, and $1/f^3$. Oscillator phase noise PSD

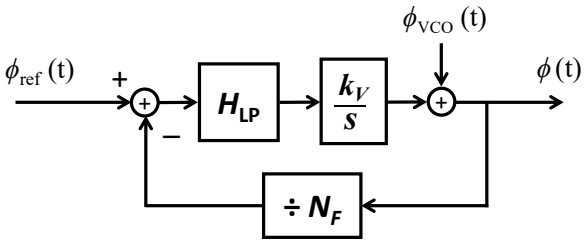


Figure 4.3: LTI system model of PLL for phase noise.

is thus described parametrically by

$$L(f) = a_0 + \frac{a_1}{f} + \frac{a_2}{f^2} + \frac{a_3}{f^3}. \tag{4.5}$$

For simplicity we assume here that all clocks are unit amplitude ($A = 1$). When a noisy oscillator is used to down-convert an RF signal, the bandpass phase noise in the oscillator output is directly transferred to the demodulated baseband signal. For digital communication this translates to rotation of the baseband symbols relative to the transmitted constellation points.

4.3.1 Phase noise in the tiled array

Fig. 4.3 shows the linearized PLL model. The signal described by this model is the *phase* of the input and output signals and therefore predicts accurately how VCO and reference phase noise are affected in the process. The phase noise at the output of this system is the sum of contributions from the reference phase and VCO phase noise; the former is identical in all tiles, whereas the latter is independent from one tile to another but identical over elements of the same tile.

The relation between phase noise at the output of the PLL and reference and VCO

phase noise is described through the model of Fig. 4.3 as

$$\tilde{\phi}(s) = \tilde{\phi}_{\text{vco}}(s) + \frac{\alpha}{s} H_{\text{LP}}(s) \left(\tilde{\phi}_{\text{ref}}(s) - \frac{\tilde{\phi}(s)}{N_f} \right)$$

where $\tilde{u}$ denotes the frequency domain representation (Laplace transform) of function u. With some manipulation, we arrive at the PLL filters applied to each phase noise source,

$$\tilde{\phi}(s) = H_{\text{ref}}^{\text{PLL}}(s)\tilde{\phi}_{\text{ref}}(s) + H_{\text{vco}}^{\text{PLL}}(s)\tilde{\phi}_{\text{vco}}(s)$$

where

$$H_{\text{ref}}^{\text{PLL}}(s) = \frac{N_f \alpha H_{\text{LP}}(s)}{N_f s + \alpha H_{\text{LP}}(s)}$$
$$H_{\text{vco}}^{\text{PLL}}(s) = \frac{N_f s}{N_f s + \alpha H_{\text{LP}}(s)}.$$

The loop filter, H_{LP}, is low pass, therefore we observe that the PLL acts as a low pass filter for reference phase noise and a high pass filter for VCO phase noise.

The filtering of reference phase noise by the PLL leaves only its *low frequency* components which results in a slow-varying signal that changes on a time scale of many symbols (hundreds or even thousands, depending on the filter bandwidth and symbol rate). Furthermore, this noise is constant over the array and passes through the linear receiver to the output where it can be tracked and compensated as described in the next section. We see, therefore, that the overall impact of reference phase noise on demodulation is very small. Of course, accounting for reference phase noise is important for channel estimation.

VCO phase noise, on the other hand, is constant for elements on one tile, but *independent over different tiles,* and therefore affects multiuser detection in a nontrivial

Table 4.1: Examples of q coefficients for nominal system.

	q_0 (Hz)	q_1	q_2 (Hz^{-1})	q_3 (Hz^{-2})
$H_{\text{vco}}^{\text{PLL}}$	5×10^9	15	2.2×10^{-6}	1.6×10^{-12}
$H_{\text{ref}}^{\text{PLL}} H_W$	1.2×10^7	0.22	6.2×10^{-8}	1.1×10^{-13}

manner. We define by $\phi_i(t)$ the VCO phase noise in the carrier of tile i and derive its variance as

$$\mathbb{E}\phi_i^2 = \sigma_\phi^2 = \int_{-B/2}^{B/2} L_{\text{vco}}(f) \left| H_{\text{vco}}^{\text{PLL}}(f) \right|^2 df \tag{4.6}$$

where $L_{\text{vco}}(f)$ is the VCO phase noise PSD and B is the system bandwidth. As we demonstrate in upcoming discussions, the impact of phase noise on the tiled multiuser system is determined by this variance, which is a linear function of the oscillator phase noise coefficients introduced in (4.5),

$$\begin{aligned} \sigma_\phi^2 &= \sum_{i=0}^{3} q_i a_i, \\ q_i &= \int_{-B/2}^{B/2} \frac{1}{f^i} \left| H_{\text{vco}}^{\text{PLL}}(f) \right|^2 df. \end{aligned} \tag{4.7}$$

After quantifying the largest tolerable phase noise variance, $\sigma_{\phi\,\max}^2$, any $L(f)$ mask that satisfies $\sum q_i a_i \leq \sigma_{\phi\,\max}^2$ maintains the desired system performance. The values of these coefficients are given for our nominal configuration in Table 4.1.

We now provide a simple upper bound (Theorem 2) on the BER for Gray coded QPSK in a SISO system, defining an equivalent SNR that combines the effects of phase noise and additive noise. We use this result in Section 4.5, where we characterize the SINR in our MIMO system, to estimate the BER.

4.3.2 Equivalent SNR for SISO BER with phase noise

Consider a unit-power QPSK modulated signal s received with Gaussian phase noise $\varphi \sim \mathcal{N}(0, \sigma_\varphi^2)$ as

$$y(t) = e^{j\varphi(t)} s(t), \quad s \in \{\frac{1}{\sqrt{2}}(\pm 1 \pm j)\}.$$

For Gray coded QPSK (which corresponds to independent BPSK streams along I and Q), the BER in the absence of additive noise is easily seen to be

$$\text{BER} = Q(\frac{\pi/4}{\sigma_\varphi}).$$

This is the same BER produced by **additive** complex Gaussian noise with variance

$$\sigma^2 = \frac{16}{\pi^2}\sigma_\varphi^2.$$

We term this variance the "equivalent *additive* noise power" for *phase* noise φ. We now show that adding this variance to that of the additive noise yields an upper bound on BER.

Theorem 2 *The BER of a Gray coded QPSK signal distorted by phase noise* $\varphi \sim \mathcal{N}(0, \sigma_\varphi^2)$ *and additive noise* $n \sim \mathcal{CN}(0, \sigma_n^2)$,

$$y = e^{j\varphi} s + n,$$

is upper bounded by the BER achieved with ***only additive*** *noise of variance*

$$\sigma^2 = \sigma_n^2 + \frac{16}{\pi^2}\sigma_\varphi^2,$$

that is,

$$BER \leq Q\left(\frac{1}{\sqrt{\sigma_n^2 + \frac{16}{\pi^2}\sigma_\varphi^2}}\right).$$

Proof. See Appendix A.

4.4 Multiuser Reception

We employ LMMSE interference suppression followed by per-user constellation tracking, as described below.

4.4.1 LMMSE reception

In the absence of phase noise, the LMMSE receiver for the model (4.1) is given by

$$\mathbf{\Gamma}_{\text{LMMSE}} = \mathbf{H}^H(\mathbf{H}\mathbf{H}^H + \sigma_\nu^2\mathbf{I})^{-1}. \tag{4.8}$$

For the model (4.4) with phase noise, we derive the approximate LMMSE receiver by treating the channel dependent distortion,

$$\mathbf{z} = \begin{bmatrix} \mathbf{H}_1 j\phi_1 \\ \mathbf{H}_2 j\phi_2 \\ \vdots \\ \mathbf{H}_{N_T} j\phi_{N_T} \end{bmatrix} \mathbf{s}$$

as an additional noise term, where $\{\phi_i\}$ are the independent VCO phase noise contributions at the output of tile PLLs.

Assuming that the user symbols are unit power and uncorrelated, $\mathbb{E}\,\mathbf{s}\mathbf{s}^H = \mathbf{I}_K$ and

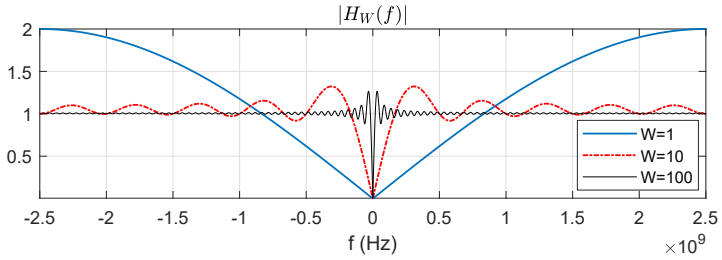


Figure 4.4: Frequency response of constellation tracking filter for different window sizes (with respect to phase signal). $W = 1$ is equivalent to differential modulation.

the covariance matrix of this distortion is of the form

$$\mathbf{C}_z = \sigma_\phi^2 \begin{bmatrix} \mathbf{H}_1\mathbf{H}_1^H & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_2\mathbf{H}_2^H & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{H}_{N_T}\mathbf{H}_{N_T}^H \end{bmatrix}$$

where σ_ϕ^2 is the filtered VCO phase noise variance described by (4.6). The optimal linear receiver for a phase noise distorted system is therefore calculated as

$$\mathbf{\Gamma}_{\text{LMMSE}} = \mathbf{H}^H(\mathbf{H}\mathbf{H}^H + \sigma_\nu^2\mathbf{I} + \mathbf{C}_z)^{-1}. \tag{4.9}$$

In terms of implementation, the naive LMMSE receiver (4.8) may be realized using one-shot channel estimates at the beginning of a duration over which the spatial channels are modeled as invariant, while the receiver (4.9) may be obtained by continuous adaptation over such an interval.

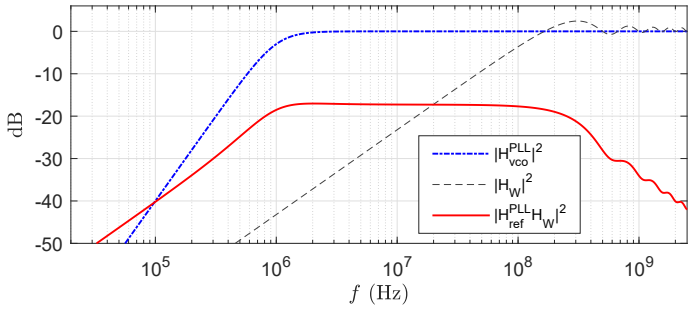


Figure 4.5: Phase noise spectrum shaping by PLL and constellation tracking filters.

4.4.2 Per-user constellation tracking

Reference phase noise is common over the array and therefore passes through the linear receiver to the output of all channels:

$$\mathbf{y} = \mathbf{\Gamma}(\mathbf{x}e^{j\phi_0}) = e^{j\phi_0}\mathbf{\Gamma}\mathbf{x}.$$

Assuming the nearest neighbor estimates are correct, i.e., $\hat{s}_k = \mathrm{NN}(y_k) = s_k$, output phase noise can be estimated after detection as

$$\hat{\phi}_{\text{out}} = \angle \frac{y_k}{\hat{s}_k}.$$

This overall phase distortion contains high pass contributions from tile VCOs and a lowpass component that is the filtered reference phase noise. By averaging the output phase noise estimate over several symbols, the low pass component is isolated and can be used to undo the reference phase drift for upcoming symbols. Assuming error-free symbol detection, this constellation tracking process can be abstracted as a windowed demeaning filter applied to the output phase noise, with impulse response

$$h_W(t) = \delta(t) - \frac{1}{W}\sum_{i=1}^{W}\delta(t - i\,T_{\text{symb}})$$

where W is the demeaning window size and $T_{\text{symb}} = 1/B$ is the symbol duration. Window size is a tunable design parameter; a smaller window can track phase drift faster and has a larger rejection bandwidth, but introduces greater noise enhancement at high frequencies. This effect is clearly shown in Fig. 4.4 where the frequency response of the tracking filter is depicted for different window sizes. High pass filtering by the constellation tracking mechanism compounded with the low pass filtering of the PLL diminishes almost all of the reference phase noise power. What remains can be formulated using the abstracted tracking model as

$$\sigma_0^2 = \int_{-B/2}^{B/2} L_{\text{ref}}(f)\left|H_{\text{ref}}^{\text{PLL}}(f)H_{\text{W}}(f)\right|^2 df \tag{4.10}$$

which depends on the filtering bandwidths of the PLL and constellation tracking process. This value is reported in Table 4.1 for the nominal configuration along with the total filtering coefficients applied to the a_i components of $L_{\text{ref}}(f)$ satisfying $\sigma_0^2 = \sum_{i=0}^{3} q_i a_i$.

The impact of drift tracking on VCO phase noise is less significant as this is already a high pass signal. This impact is slight noise enhancement at the higher end of the spectrum (evident in Fig. 4.4) which, in our running example, increases output VCO phase noise by about 0.7%. If the bandwidth of H_{vco}^{PLL} is smaller than that of H_W, drift tracking may *decrease* the output VCO phase noise covariance. It is worth noting that constellation tracking has no effect on the *cross-user* interference caused by VCO phase noise since, (a) it is applied *after* multiuser reception where crosstalk is produced, and (b) phase alteration does not alter the distribution of this interference as it is a zero-mean

complex Gaussian variable.

4.5 Interference Analysis

For the signal model (4.4), the output of a linear receiver $\mathbf{\Gamma}$ is approximated by

$$\mathbf{y} = \mathbf{\Gamma H s} + \mathbf{\Gamma}\boldsymbol{\nu} + \sum_{i=1}^{N_T} j\phi_i \mathbf{\Gamma}_i \mathbf{H}_i \mathbf{s}$$

where $\mathbf{\Gamma}_i$ denotes the i'th $K \times N_0$ block of $\mathbf{\Gamma}$. That is,

$$\mathbf{\Gamma} = [\mathbf{\Gamma}_1 \, \mathbf{\Gamma}_2 \, ... \, \mathbf{\Gamma}_{N_T}].$$

The naive LMMSE receiver (4.8) focuses on suppression of the coherent (across tiles) interference in the first term. For our nominal load factor of $\beta = \frac{1}{4}$, and under the assumed minimum spatial frequency separation of $2\pi/N$, the resulting LMMSE receiver does not lead to a significant attenuation of the desired signal. The LMMSE receiver (4.9) does much the same, since the phase noise causing the third term is significantly smaller (by design) than the coherent interference. Furthermore, the third term is not coherent across tiles ($\{\phi_i\}$ are independent), and tile-level interference suppression is not feasible in the scaling regime of interest to us, where the number of users is greater than the (fixed) number of antennas per tile. We estimate tile-level interference, therefore, under the following approximation.

Approximation: *At the tile-level, both LMMSE variants are assumed to be aligned with the spatial matched filters for the users:*

$$\mathbf{\Gamma}_i \approx \frac{1}{N} \mathbf{H}_i^H. \tag{4.11}$$

We have verified this approximation via extensive simulations, but report only an example result: at SNR of 14 dB, the normalized correlation at the tile-level between the LMMSE receiver and the spatial matched filter, at the tile level, is found to exceed 0.97 with probability 99%.

Under the approximation (4.11), we obtain (see Appendix B) that the diagonal entries are equal to

$$(\mathbf{\Gamma}_i \mathbf{H}_i)_{k,k} = \frac{N_0}{N} \tag{4.12}$$

and the off-diagonal entries are zero-mean with variance

$$\mathbb{E}\left|(\mathbf{\Gamma}_i \mathbf{H}_i)_{k,l}\right|^2 = \frac{N_0}{N^2}, \qquad (k \neq l) \tag{4.13}$$

We can now state the following result.

Theorem 3 *Under the approximation (4.11), phase noise causes multiplicative self-noise and additive cross-user interference described by*

$$y_k \approx \gamma e^{j\psi} s_k + I \tag{4.14}$$

where

$$\begin{aligned} &\mathbb{E}\psi^2 = \frac{\sigma_\phi^2}{N_T}, \qquad \mathbb{E}\gamma = 1 - \frac{1}{2}\sigma_\phi^2, \\ &\mathbb{E}|I|^2 = \frac{K-1}{N}\sigma_\phi^2 \approx \beta\sigma_\phi^2. \end{aligned} \tag{4.15}$$

Proof. Setting aside the additive contribution of thermal noise,

$$\mathbf{y} = \mathbf{\Gamma} \begin{bmatrix} \mathbf{H}_1 e^{j\phi_1} \\ \mathbf{H}_2 e^{j\phi_2} \\ \vdots \\ \mathbf{H}_{N_T} e^{j\phi_{N_T}} \end{bmatrix} \mathbf{s}.$$

The output of channel k is a combination of contributions from the desired and interfering users which we separate as

$$y_k = \left(\sum_{i=1}^{N_T} (\mathbf{\Gamma}_i \mathbf{H}_i)_{k,k} e^{j\phi_i} \right) s_k + \sum_{l \neq k} \left(\sum_{i=1}^{N_T} (\mathbf{\Gamma}_i \mathbf{H}_i)_{k,l} e^{j\phi_i} \right) s_l.$$

The multiplicative self-noise is derived using the second order Taylor expansions of sine and cosine functions:

$$\begin{aligned} \sum_{i=1}^{N_T} (\mathbf{\Gamma}_i \mathbf{H}_i)_{k,k} e^{j\phi_i} &\approx \frac{N_0}{N} \sum_{i=1}^{N_T} (\cos\phi_i + j \sin\phi_i) \approx \frac{1}{N_T} \sum_{i=1}^{N_T} (1 - \frac{1}{2}\phi_i^2) + j\phi_i \\ &\approx \left(1 - \frac{1}{2N_T} \sum_{i=1}^{N_T} \phi_i^2 \right) + j \left(\frac{1}{N_T} \sum_{i=1}^{N_T} \phi_i \right) = \gamma + j\psi \approx \gamma e^{j\psi} \end{aligned}$$

where $\gamma = 1 - \frac{1}{2N_T} \sum_{i=1}^{N_T} \phi_i^2$ is an average of identically distributed non-zero-mean variables, and is therefore well-approximated by its expected value,

$$\gamma \approx \mathbb{E}\gamma = 1 - \frac{\sigma_\phi^2}{2}$$

and $\psi = \frac{1}{N_T} \sum_{i=1}^{N_T} \phi_i$ is an average of zero-mean i.i.d. Gaussian random variables, with variance

$$\mathbb{E}\psi^2 = \frac{\sigma_\phi^2}{N_T}.$$

We evaluate cross-user interference by using the small-phase approximation (first order Taylor expansion) to obtain

$$\sum_{i=1}^{N_T}(\mathbf{\Gamma}_i\mathbf{H}_i)_{k,l}\, e^{j\phi_i} = \sum_{i=1}^{N_T}(\mathbf{\Gamma}_i\mathbf{H}_i)_{k,l} + \sum_{i=1}^{N_T} j\phi_i(\mathbf{\Gamma}_i\mathbf{H}_i)_{k,l}$$
$$= (\mathbf{\Gamma H})_{k,l} + \sum_{i=1}^{N_T} j\phi_i(\mathbf{\Gamma}_i\mathbf{H}_i)_{k,l} \approx \sum_{i=1}^{N_T} j\phi_i(\mathbf{\Gamma}_i\mathbf{H}_i)_{k,l}$$

with $(\mathbf{\Gamma H})_{k,l} \approx 0$ resulting from the LMMSE receiver effectively suppressing cross terms in $\mathbf{\Gamma Hs}$ using all N degrees of freedom. The covariance of cross-user interference is thus found to be

$$\mathbb{E}|I|^2 = \mathbb{E}\left|\sum_{l\neq k}\left(\sum_{i=1}^{N_T}(\mathbf{\Gamma}_i\mathbf{H}_i)_{k,l}e^{j\phi_i}\right)s_l\right|^2$$
$$= (K-1)\, N_T\, \sigma_\phi^2\, \mathbb{E}|(\mathbf{\Gamma}_i\mathbf{H}_i)_{k,l}|^2$$
$$= (K-1)\,\frac{N_T N_0}{N^2}\,\sigma_\phi^2 = \frac{K-1}{N}\sigma_\phi^2 \lesssim \beta\sigma_\phi^2$$

□

Accounting for all multiplicative and additive noise terms in the system, we arrive at the following model for the output signal:

$$y_k = \gamma e^{j(\psi+\phi_0)}s_k + I + \nu'$$

where

$$\gamma = 1 - \frac{\sigma_\phi^2}{2}, \qquad \mathbb{E}\psi^2 = \frac{\sigma_\phi^2}{N_T}, \qquad \mathbb{E}\phi_0^2 = \sigma_0^2,$$

$$\mathbb{E}I^2 = \beta\sigma_\phi^2, \qquad \mathbb{E}|\nu'|^2 = \frac{\sigma_\nu^2}{N}.$$

Fig. 4.6 shows scatter plots of the received QPSK symbols in the I-Q plane for different

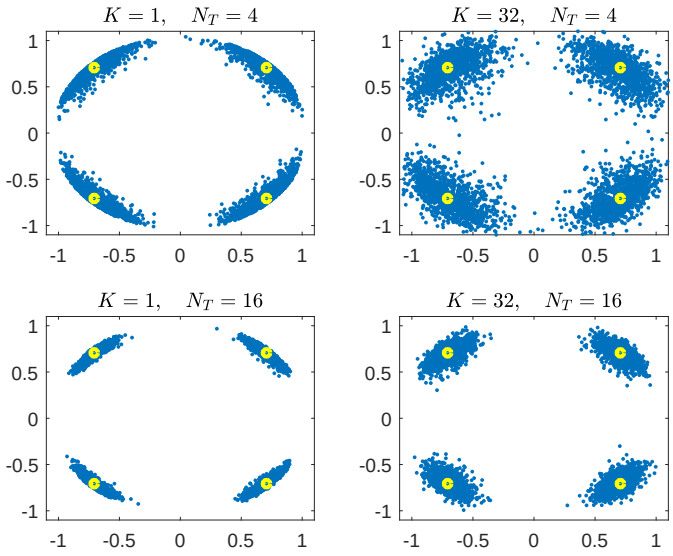


Figure 4.6: Scatter plot of received QPSK symbols on a 256-element array for different load factors and number of tiles. The output phase noise decreases as number of tiles increases and interference is proportional to number of users. Additive noise has been set to zero to emphasize these effects.

values of K and N_T. Based on the preceding analysis, if load factor and tile size are kept constant, phase noise is not a bottleneck in scaling to larger arrays. The cross-user interference caused by phase noise only depends on the ratio of users to array size and is therefore constant, while output phase noise variance *decreases* as the number of tiles grows. In fact, we expect performance to *improve* as the system is scaled up, as long as loading, tile size, oscillator PSD, and beamformed SNR (σ_ν^2/N) are fixed.

Predicting the BER of a phase-distorted system. Using Theorem 2, we can determine the equivalent noise variance for our tiled system as

$$\sigma_{\text{eq}}^2 = \frac{1}{(1-\frac{1}{2}\sigma_\phi^2)^2}\left(\frac{\sigma_\nu^2}{N} + \beta\sigma_\phi^2 + \frac{16}{\pi^2}\left(\frac{\sigma_\phi^2}{N_T} + \sigma_0^2\right)\right)$$

This value is an expectation and actual SINR varies across users. Since we assume power leveling, this variation is primarily due to difference in cross-user interference. In order

to provide a pessimistic prediction, we substitute the average cross-talk power $\beta\sigma_\phi^2$ with 3 standard deviations above its mean,

$$\left(\beta + 3\,\frac{0.82\sqrt{\beta}}{N_T}\right)\sigma_\phi^2$$

(derivation of the variance of the cross-talk power omitted due to space considerations). Since BER is dominated by the worst-case, we expect this pessimistic approach to be accurate. The equivalent SINR is thus modified to

$$\sigma_{\text{eq}}^2 = \frac{1}{(1-\frac{1}{2}\sigma_\phi^2)^2}\left(\frac{\sigma_\nu^2}{N} + (\beta + 2.46\frac{\sqrt{\beta}}{N_T})\sigma_\phi^2 + \frac{16}{\pi^2}\left(\frac{\sigma_\phi^2}{N_T} + \sigma_0^2\right)\right). \tag{4.16}$$

We now account for the reduction in SINR due to the reduction in signal power (or equivalently, noise enhancement) caused by suppression of the coherent interference. At moderate load factors (e.g., $\beta = 1/4$) and SNRs, this SINR penalty is well approximated by that due to a zero-forcing receiver. Let ρ denote the normalized cross-correlation between the spatial channels for two randomly chosen users. If the spatial frequencies are uniform over $(-\pi, \pi)$, then $E[|\rho|^2] = \frac{1}{N}$ (see Appendix B). When we enforce a minimum spatial frequency separation of $2\pi/N$, we can actually show that $E[|\rho|^2] \leq \frac{0.1}{N}$ (again, see Appendix B). In order to evaluate the impact of scaling, therefore, we set

$$E[|\rho|^2] = \frac{\alpha}{N} \tag{4.17}$$

Now, denoting by ρ_{lk} the normalized cross-correlation between the spatial responses for users l and k, the signal power S relative to matched filtering, is bounded as follows:

$$S \geq 1 - \sum_{l \neq k} |\rho_{lk}|^2 \to 1 - (K-1)E[|\rho|^2] \geq 1 - \alpha\beta \tag{4.18}$$

where the limit is as $K, N \to \infty$ with β fixed, and where we have used (4.17). Using the right-hand side of (4.18) as an approximation, we obtain the following pessimistic prediction of the equivalent SINR

$$\mathrm{SINR}_{\mathrm{eq}} \approx \frac{1-\alpha\beta}{\sigma_{\mathrm{eq}}^2}. \tag{4.19}$$

which in turn yields a pessimistic approximation for BER [96]:

$$\mathrm{BER} \approx Q\left(\sqrt{\mathrm{SINR}_{\mathrm{eq}}}\right). \tag{4.20}$$

This prediction, with $\alpha = 0.1$ in (4.19), is expected to be accurate in our regime of interest of $K > N_0$ and moderate SNR. In the underloaded regime, per-tile interference suppression becomes feasible and, at high enough SNR, partial per-tile interference suppression is incentivized even for $K > N_0$. In such settings (not of interest in our scaling regime), our prediction is still pessimistic, but is not as good of an approximation; i.e., performance can be significantly better than predicted.

In the next section we summarize our design framework, and provide simulation results to validate our analytical predictions.

4.6 Numerical results

The models presented here can be used to provide an **analytical cross-layer design framework** using four key observations.

- The reference oscillator phase noise is filtered by the PLL and constellation tracking filters and appears at the output of all channels with variance $\sigma_0^2 = \int L_{\mathrm{ref}}(f) \left| H_{\mathrm{ref}}^{\mathrm{PLL}}(f) H_W(f) \right|^2 df$.

- VCO noise is filtered by the PLL resulting in tile phase noise with variance

$$\sigma_\phi^2 = \int L_{\mathrm{VCO}}(f) \left| H_{\mathrm{vco}}^{\mathrm{PLL}}(f) \right|^2 df.$$

  At the receiver output this produces phase noise of variance σ_ϕ^2/N_T and additive interference of power $\beta\sigma_\phi^2$.

- System performance is determined by parameters σ_0^2, σ_ϕ^2, N_T, β, and σ_ν^2 (beamformed SNR). A pessimistic prediction is given by the equivalent SINR

$$\mathrm{SINR}_{\mathrm{eq}} = \frac{1-\alpha\beta}{\sigma_{\mathrm{eq}}^2},$$

$$\sigma_{\mathrm{eq}}^2 = (1-\sigma_\phi^2)^{-2}\left(\mathrm{SNR}^{-1} + (\beta + \frac{2.46\sqrt{\beta}}{N_T})\sigma_\phi^2 + \frac{16}{\pi^2}\left(\frac{\sigma_\phi^2}{N_T} + \sigma_0^2 \right) \right),$$

  where $\alpha = 0.1$ for the specifications of our system model.

- Permissible phase noise PSD can be expressed as a *linear constraint* on the $L(f)$ coefficients,

$$\sum q_i a_i \leq \sigma_\phi^2 \text{ or } \sigma_0^2,$$

  with q factors obtained by (4.7).

Using the preceding guidelines, trade-offs between different design choices are modeled as simple analytical relationships that predict system performance with reasonable accuracy. In this section we provide numerical validation for the above framework and examine the scaling laws derived from it.

We first provide an example of an acceptable **phase noise mask** for our nominal system specifications. We use the model of (4.5) to generate phase noise signals. The reference phase noise PSD is set lower than that of the VCO by a factor of N_f. In

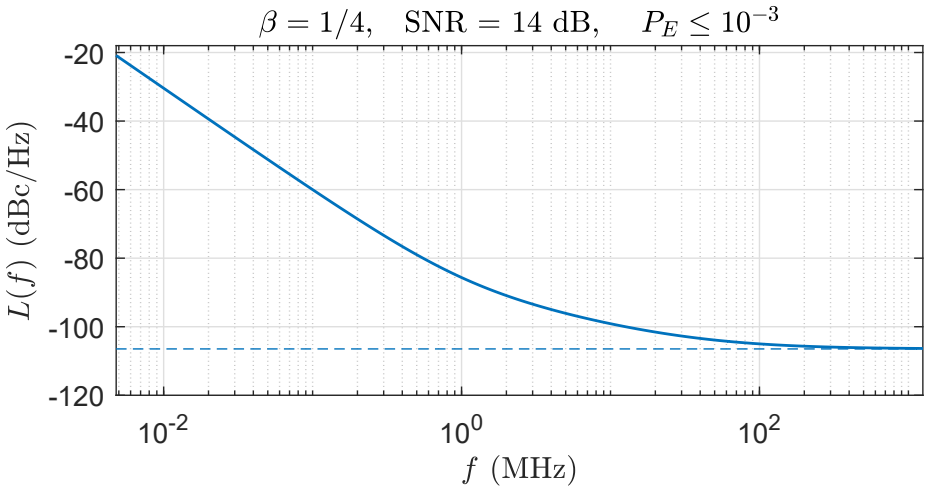


Figure 4.7: Acceptable phase noise mask for target BER 10^{-3} and 10 Gbps rate in nominal system. Parameters: $a_0 = 2.25 \times 10^{-11}$ W/Hz, $a_1 = 9 \times 10^{-4}$ W, $a_2 = 9 \times 10^{2}$ WHz, $a_3 = 9 \times 10^{8}$ WHz^2.

practice, the reference is likely to be a high quality crystal resonance oscillator with very low phase noise. The shape of the curve (a_i parameters) is chosen such that low-pass components have approximately the same combined impact as the constant component for our nominal system. Fig. 4.7 shows the resulting $L(f)$ mask which indicates feasible phase noise requirements for our target system, as THz oscillators with lower $L(f)$ have been reported in the literature [97].

Using the PSD curve shown in Fig. 4.7, we isolate the two distortion terms described in Theorem 3 as follows. To measure **self-noise**, we set up the end-to-end multiuser system with $\beta = 1/4$ and LMMSE reception (as described in Section 4.4). We then set the transmitted symbol, $s_k(t)$, to zero for all but one user throughout each simulation sequence and set the additive noise to zero (but use the nominal value for σ_ν^2 in deriving the LMMSE receiver). Fig. 4.8 depicts self-noise for the designated user as a function of tile size and phase noise variance for a 256-element array. The analytical predictions are also plotted for comparison. We observe that simulation results follow the analytical prediction closely.

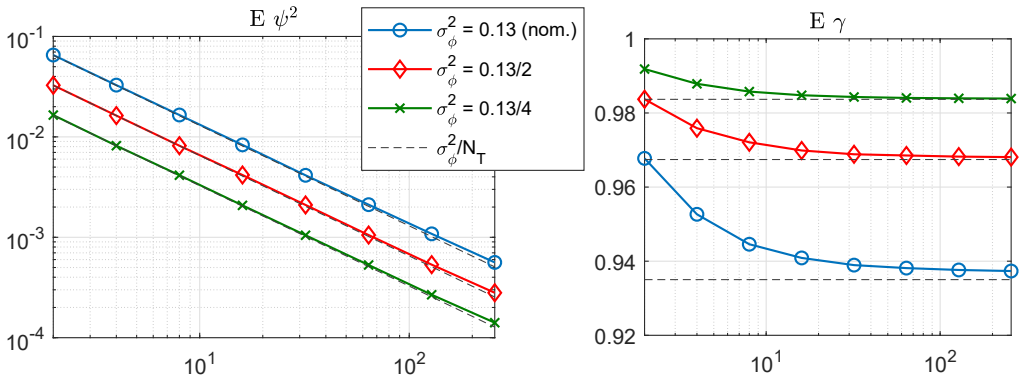


Figure 4.8: Self interference phase noise and amplitude attenuation for 256 element array with 64 users.

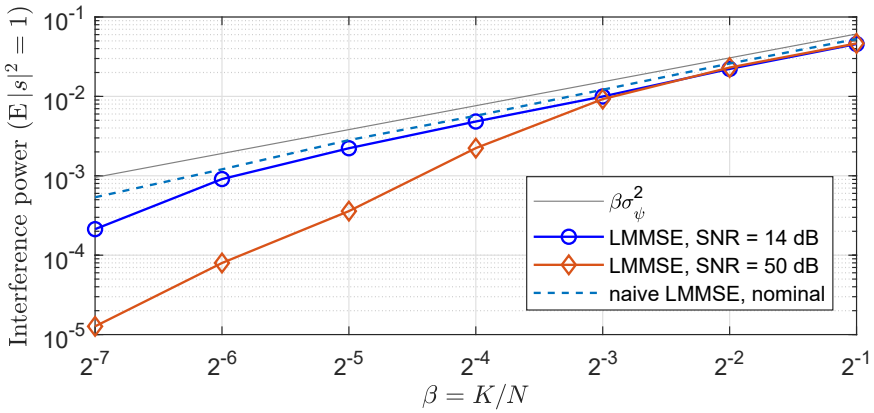
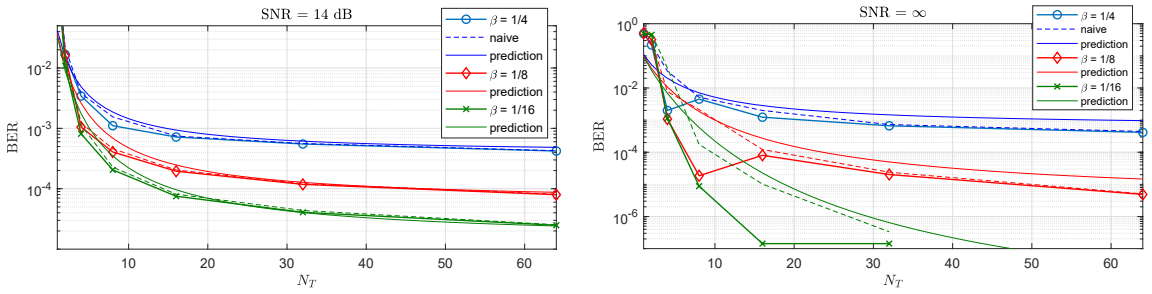


Figure 4.9: Scaling of cross-talk with load factor.

To isolate the **cross-user interference**, we perform a similar simulation but this time set the designated user's signal to zero while all other users remain active. This way we only get cross-user interference signal at the designated output, the variance of which is the interference power. In Fig. 4.9 we report the cross-user interference variance averaged over many realizations for naive LMMSE (that ignores the effect of phase noise) and optimal LMMSE for nominal and high SNR. Both receivers follow the analytical prediction, $\mathbb{E}|I|^2 = \beta\sigma_\phi^2$, fairly closely in most regimes. For an underloaded system at high SNR, analytical results overestimate interference for LMMSE reception since *per tile* interference suppression becomes possible, as discussed in Section 4.5.



(a) (b)

Figure 4.10: Full system simulation results for the nominal configuration using $L(f)$ function of Figure 4.7 varying load factor and number of tiles ($N = 256$ fixed). Optimal LMMSE, naive LMMSE (ignoring phase noise), and prediction of (4.16) and (4.19) plotted for comparison.

Full system simulations that include the effects of VCO and reference phase noise, as well as realistic implementation of constellation tracking were performed to quantify the effect of system parameters on performance. Fig. 4.10a depicts performance as a function of load factor and tile size, assuming beamformed SNR of 14 dB and phase noise PSD depicted in Fig. 4.7. When the array is divided into a larger number of tiles, self-noise is suppressed and BER decreases, but if β remains constant the cross-user interference does not change with N_T and therefore creates a performance floor. In Fig. 4.10b we depict the same results for a high SNR system. As expected, we see that BER is overestimated considerably for optimal LMMSE reception in the underloaded regime.

Finally, Fig. 4.11 shows how performance scales with beamformed link SNR in the presence and absence of phase noise. The performance floor caused by phase noise, especially its cross-user interference effect, is clearly visible in this figure. This floor can only be suppressed by reducing the load factor as shown in Fig. 4.10. By comparing realistic drift tracking with the idealized model of (4.10) we see that, at low SINR, a decoding error can deteriorate tracking performance and cause additional errors in consecutive symbols. This effect can be mitigated by adding a margin to the target SINR. For our nominal setup, idealized performance is recovered with 0.5 dB higher

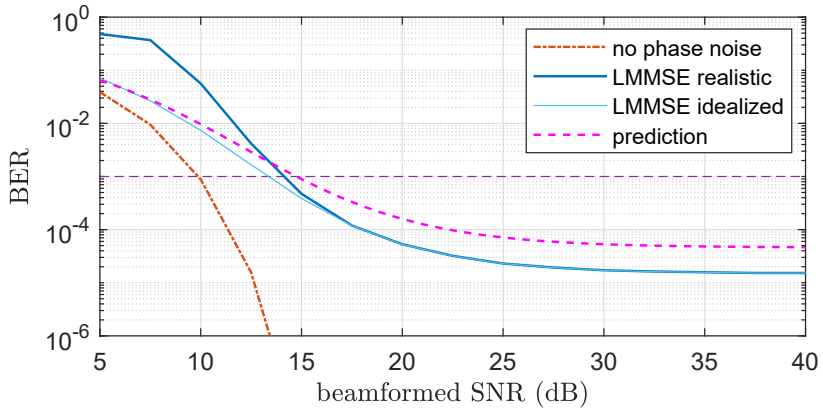


Figure 4.11: Performance (BER) as a function of SNR with and without phase noise. Solid narrow curve (light blue) assuming idealized constellation tracking, i.e., no error propagation.

SINR, which is achieved by scaling down all noise terms uniformly, i.e., increasing SNR and reducing $L(f)$ each by 0.5 dB. As expected, our analytical predictions are only pessimistic for the idealized case and may underestimate BER for a realistic system at low SINR.

Chapter 5

Conclusions and Future Work

In the preceding chapters we have outlined a roadmap to next generation mobile networks through exploiting the advantages of mmWave frequencies, and addressed some of the challenges in this path.

We identified point-to-point mmWave links as a viable option for high speed wireless backhauling of picocells, enabling shrinking of cell sizes and orders of magnitude increase in frequency reuse. Possible future directions of research also include speeding up the scheduling procedure using iterative optimization approaches or dynamic programming, possibly at the cost of slightly suboptimal solutions. Distributed scheduling algorithms are also an attractive option that pave the way for self-organizing networks that do not rely on centralized control. Jointly optimizing the placement of nodes and neighbor associations can also improve backhaul performance and is an interesting avenue of research.

High speed communication with mobile users inside the picocells is facilitated by deploying highly directional mmWave arrays that reduce interference and sustain the SNR required for Gbps signaling to users. We proposed a noncoherent compressive channel estimation framework that provides low-overhead yet accurate tracking of mobile users in the absence of all-digital frontends. While the single path detection scheme was

evaluated on a testbed, experimental validation of the multipath channel estimation algorithm is an important next step. Accounting for the *receive* antenna array is also left to future work.

One immediate benefit of the picocellular network is the possibility of enhanced localization accuracy without relying on satellite GPS. With a rich enough deployment of access points, fine grained positioning of a mobile device is possible by compressive estimation of its angle toward several base stations in its vicinity. In fact, such a direction-based localization scheme can be used as a building block for many applications, such as calibration of distributed sensing platforms, e.g., a herd of drones cooperating to form an ad hoc radar array can use pairwise direction measurements to map out their relative locations allowing them to operate as a distributed array.

In this work, we also demonstrated the scalability of high-bandwidth multi-user massive MIMO in the presence of phase noise with modest hardware requirements and proposed a tractable design framework based on our analysis. While the present work is a useful first step in cross-layer design of MU-MIMO systems, further investigation into the impact of phase noise on channel estimation (CSI), and the compounded effects of CSI error, ADC, receiver nonlinearity, and phase noise is a necessary next step, as is the experimental validation of present and future results. Furthermore, as the system grows large, digital processing can become a bottleneck for LMMSE multi-user reception and scalable redesign of backend processing is necessary. Algorithms that rely on localized beamspace information are a promising approach for speeding up array processing.

MmWave picocells and the smart city

As mmWave picocells come closer to reality, a new landscape emerges with countless applications and unforeseen challenges, some of which are depicted in Figure 5.1.

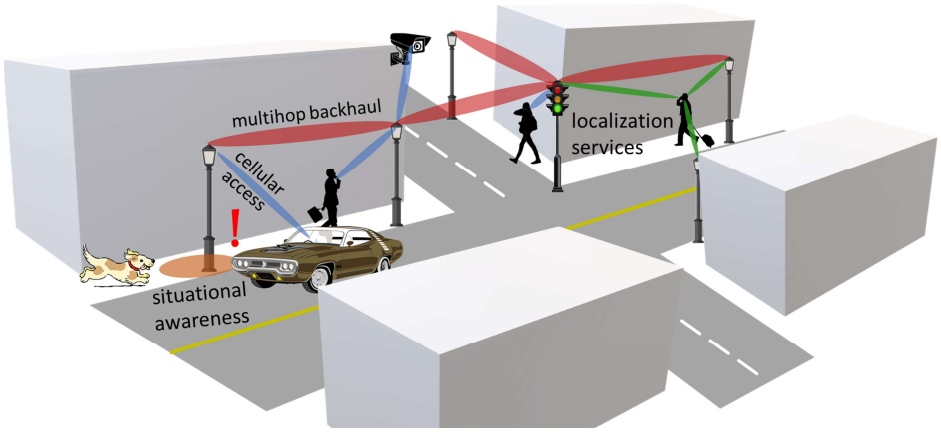


Figure 5.1: Emerging applications of mmWave and THz in the wireless landscape.

With wireless backhauling, compact and easy to deploy access points, and virtually interference-free directional links to users, the picocellular network has the potential to become a robust, self-organizing infrastructure, that can be expanded easily in a plug-and-play fashion. The capabilities of such a system are well beyond cellular mobile service, and can power the migration to smart cities, supporting real-time tasks such as traffic control, autonomous vehicles, surveillance, localization, and Internet of things. Furthermore, large portions of data processing for these applications can be incorporated into the network, similar to distributed cloud-based processing, with the mobile infrastructure evolving into the "brain of the city". This vision opens up a myriad of open problems and interesting challenges from robust management and control protocols to security and privacy.

Appendix A

Proof of Theorem 2

Gray coded QPSK corresponds to independent bits sent along I and Q. By symmetry, condition without loss of generality on transmitting $s = e^{j\pi/4}$, and on the probability of error in the I bit. Consider the two signals

$$y_1 = se^{j\varphi} + n$$

$$y_2 = s + n_\varphi + n$$

where $\varphi \sim \mathcal{N}(0, \sigma_\phi^2)$ and $n_\varphi \sim \mathcal{CN}(0, \frac{16}{\pi^2}\sigma_\phi^2)$, meaning real and imaginary parts of n_φ are real zero-mean Gaussian variables with variance $\frac{8}{\pi^2}\sigma_\phi^2$.

An error in the I bit occurs when y crosses the vertical boundary into the region $\mathfrak{Re}(y) < 0$. In the first signal model, this error occurs when

$$\mathfrak{Re}(y_1) = \cos(\pi/4 + \varphi) + n_i < 0$$

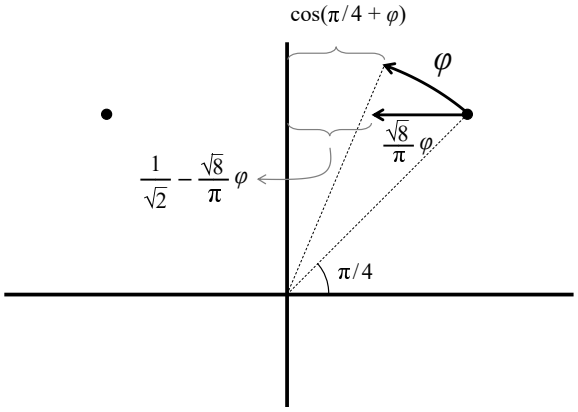


Figure A.1: Margin left for additive noise n after phase distortion and equivalent additive distortion.

and in the second model when

$$\mathfrak{Re}(y_2) = \frac{1}{\sqrt{2}} + \mathfrak{Re}(n_\varphi) + n_i < 0$$

where n_i is the real part of the complex Gaussian variable n. To compare the probability of these two occurrences we define the additive variable as

$$\mathfrak{Re}(n_\varphi) = -\frac{\sqrt{8}}{\pi}\varphi$$

which is a zero-mean Gaussian that satisfies $\mathbb{E}\left(\mathfrak{Re}(n_\varphi)\right)^2 = \frac{8}{\pi^2}\sigma_\varphi^2$. The horizontal error margin left after the effect of φ and n_φ is, respectively,

$$m_1 = \cos(\pi/4 + \varphi), \qquad m_2 = \frac{1}{\sqrt{2}} - \frac{\sqrt{8}}{\pi}\varphi,$$

as depicted in Fig. A.1. For any $\varphi \in (0, \pi/4)$ we have $m_1 - m_2 \geq 0$ because $m_1 - m_2$ is

a convex function of φ in this domain with value 0 at the boundaries,

$$
\begin{aligned}
&@\,\varphi = 0: \qquad m_1 = m_2 = 1/\sqrt{2} \\
&@\,\varphi = \pi/4: \quad m_1 = m_2 = 0 \\
&\frac{\partial^2}{\partial \varphi^2}(m_1 - m_2) = -\cos(\pi/4 + \varphi) < 0,\ \forall \varphi \in (0, \pi/4).
\end{aligned}
$$

We therefore conclude that

$$\mathrm{BER} = P[\text{I bit wrong}] = \Pr(\mathfrak{Re}(y_1) < 0) \leq \Pr(\mathfrak{Re}(y_2) < 0).$$

□

Appendix B

Computations of spatial inner products

Let $\mathbf{a}(\omega) = (1, e^{j\omega}, ..., e^{j(n-1)\omega})^T$ denote the response of an n-element linear array to spatial frequency ω. Note that

$$||\mathbf{a}(\omega)||^2 = n \tag{B.1}$$

Applying this for $n = N_0$ gives us (4.12). The magnitude of the inner product between the responses for two different spatial frequencies ω_1 and ω_2, with $\Delta\omega = \omega_1 - \omega_2$, is given by

$$|\langle \mathbf{a}(\omega_1), \mathbf{a}(\omega_2) \rangle| = |1 + e^{j\Delta\omega} + ... + e^{j(n-1)\Delta\omega}| = \left| \frac{\sin n\Delta\omega/2}{\sin \Delta\omega/2} \right| \tag{B.2}$$

The corresponding normalized inner product is the magnitude of the well-known Dirichlet kernel:

$$\kappa_n(\Delta\omega) = \left| \frac{\sin n\Delta\omega/2}{n \sin \Delta\omega/2} \right| \tag{B.3}$$

If ω_1, ω_2 are independent and uniform over $(-\pi, \pi)$, then modulo 2π, $\Delta\omega$ is uniform over $(-\pi, \pi)$. Squaring the $|\cdot|$ term in (B.2) and taking expectations, the contribution of

cross-terms of the form $e^{jk\Delta\omega}$, where k is a nonzero integer, is zero. We therefore obtain

$$E\left[|\langle \mathbf{a}(\omega_1), \mathbf{a}(\omega_2)\rangle|^2\right] = n \tag{B.4}$$

Applying this for $n = N_0$ gives us (4.13). We may also write this as

$$E\left[|\kappa_n(\Delta\omega)|^2\right] = \frac{1}{n} \tag{B.5}$$

Applying this for $n = N$ gives us $a = 1$ in (4.17).

To see what happens when we enforce a minimum spatial separation $2\pi/n$, note that $|x| \geq |\sin x|$, so that $\kappa_n(x) \leq |\frac{\sin(nx/2)}{nx/2}|$. We can now calculate that

$$\begin{aligned}\frac{1}{2\pi}\int_{-2\pi/n}^{2\pi/n} \kappa_n^2(x)dx &\geq \frac{1}{2\pi}\int_{-2\pi/n}^{2\pi/n} \left|\frac{\sin(nx/2)}{nx/2}\right|^2 dx \\ &= \frac{1}{n\pi}\int_{-\pi}^{\pi} \left|\frac{\sin(u)}{u}\right|^2 du \geq \frac{0.90}{n}.\end{aligned}$$

Setting $n = N$, under the assumed minimum spatial separation of $2\pi/N$, we therefore obtain that

$$\mathbb{E}|\rho|^2 \leq \frac{0.1}{N}$$

so that we may set $\alpha = 0.1$ in (4.17).

Bibliography

[1] X. Ge, H. Cheng, M. Guizani, and T. Han, *5G wireless backhaul networks: challenges and research advances*, *IEEE Network* **28** (2014), no. 6 6–11.

[2] Y. Zhu, Z. Zhang, Z. Marzi, C. Nelson, U. Madhow, B. Y. Zhao, and H. Zheng, *Demystifying 60GHz outdoor picocells*, in *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking*, MobiCom '14, (New York, NY, USA), pp. 5–16, ACM, 2014.

[3] Z. Marzi, U. Madhow, and H. Zheng, *Interference analysis for mm-wave picocells*, in *Global Communications Conference (GLOBECOM), 2015 IEEE*, pp. 1–6, IEEE, 2015.

[4] S. Singh, M. N. Kulkarni, A. Ghosh, and J. G. Andrews, *Tractable model for rate in self-backhauled millimeter wave cellular networks*, *IEEE Journal on Selected Areas in Communications* **33** (2015), no. 10 2196–2211.

[5] K. Zheng, L. Zhao, J. Mei, M. Dohler, W. Xiang, and Y. Peng, *10 Gb/s hetsnets with millimeter-wave communications: access and networking-challenges and protocols*, *IEEE Communications Magazine* **53** (2015), no. 1 222–231.

[6] R. Taori and A. Sridharan, *Point-to-multipoint in-band mmwave backhaul for 5G networks*, *IEEE Communications Magazine* **53** (2015), no. 1 195–201.

[7] M. E. Rasekh, D. Guo, and U. Madhow, *Interference-aware routing and spectrum allocation for millimeter wave backhaul in urban picocells*, in *Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on*, pp. 1–7, IEEE, 2015.

[8] G. Meurant, *Handbook of convex geometry*. Elsevier, 2014.

[9] M. E. Rasekh, D. Guo, and U. Madhow, *Joint routing and resource allocation for millimeter wave picocellular backhaul*, *IEEE Transactions on Wireless Communications* (2019).

[10] D. Ramasamy, S. Venkateswaran, and U. Madhow, *Compressive adaptation of large steerable arrays*, in *Information Theory and Applications Workshop (ITA), 2012*, pp. 234–239, IEEE, 2012.

[11] D. Ramasamy, S. Venkateswaran, and U. Madhow, *Compressive tracking with 1000-element arrays: A framework for multi-Gbps mm wave cellular downlinks*, in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 690–697, Oct, 2012.

[12] Z. Marzi, D. Ramasamy, and U. Madhow, *Compressive channel estimation and tracking for large arrays in mm-wave picocells*, *IEEE Journal of Selected Topics in Signal Processing* **10** (2016), no. 3 514–527.

[13] M. E. Rasekh, Z. Marzi, Y. Zhu, U. Madhow, and H. Zheng, *Noncoherent mmWave path tracking*, in *Proceedings of the 18th International Workshop on Mobile Computing Systems and Applications*, pp. 13–18, ACM, 2017.

[14] "Facebook Terragraph and Project ARIES." https://code.facebook.com/posts/1072680049445290/introducing-facebook-s-new-terrestrial-connectivity-systems-terragraph-and-project-aries/.

[15] M. E. Rasekh and U. Madhow, *Noncoherent compressive channel estimation for mm-wave massive MIMO*, in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, pp. 889–894, IEEE, 2018.

[16] M. E. Rasekh, M. Abdelghany, U. Madhow, and M. Rodwell, *Phase noise analysis for mmwave massive MIMO: a design framework for scaling via tiled architectures*, in *2019 53rd Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, IEEE, 2019.

[17] M. E. Rasekh, M. Abdelghany, U. Madhow, and M. Rodwell, *Phase noise in modular millimeter wave massive MIMO*, *arXiv preprint arXiv:1910.09095* (2019).

[18] Y. Yi, G. De Veciana, and S. Shakkottai, *On optimal MAC scheduling with physical interference*, in *IEEE INFOCOM 2007-26th IEEE International Conference on Computer Communications*, pp. 294–302, IEEE, 2007.

[19] T. ElBatt and A. Ephremides, *Joint scheduling and power control for wireless ad hoc networks*, *IEEE Transactions on Wireless communications* **3** (2004), no. 1 74–85.

[20] P. Gupta and P. R. Kumar, *The capacity of wireless networks*, *IEEE Transactions on information theory* **46** (2000), no. 2 388–404.

[21] D. Yuan, H.-Y. Lin, J. Widmer, and M. Hollick, *Optimal joint routing and scheduling in millimeter-wave cellular networks*, in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pp. 1205–1213, IEEE, 2018.

[22] G. Narlikar, G. Wilfong, and L. Zhang, *Designing multihop wireless backhaul networks with delay guarantees*, *Wireless Networks* **16** (2010), no. 1 237–254.

[23] X. Lin and S. Rasool, *A distributed joint channel-assignment, scheduling and routing algorithm for multi-channel ad-hoc wireless networks*, in *IEEE INFOCOM 2007-26th IEEE International Conference on Computer Communications*, pp. 1118–1126, IEEE, 2007.

[24] C.-Y. Hong and A.-C. Pang, *3-approximation algorithm for joint routing and link scheduling in wireless relay networks*, *IEEE Transactions on Wireless Communications* **8** (2009), no. 2 856–861.

[25] X. Wang and J. Garcia-Luna-Aceves, *Embracing interference in ad hoc networks using joint routing and scheduling with multiple packet reception*, *Ad Hoc Networks* **7** (2009), no. 2 460–471.

[26] J. Tang, G. Xue, C. Chandler, and W. Zhang, *Interference-aware routing in multihop wireless networks using directional antennas*, in *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, vol. 1, pp. 751–760, IEEE, 2005.

[27] B. Mumey, J. Tang, and T. Hahn, *Joint stream control and scheduling in multihop wireless networks with MIMO links*, in *2008 IEEE International Conference on Communications*, pp. 2921–2925, IEEE, 2008.

[28] J. Zhang, H. Wu, Q. Zhang, and B. Li, *Joint routing and scheduling in multi-radio multi-channel multi-hop wireless networks*, in *2nd International Conference on Broadband Networks, 2005.*, pp. 631–640, IEEE, 2005.

[29] R. Bhatia and M. Kodialam, *On power efficient communication over multi-hop wireless networks: Joint routing, scheduling and power control*, in *IEEE INFOCOM 2004*, vol. 2, pp. 1457–1466, IEEE, 2004.

[30] M. Alicherry, R. Bhatia, and L. E. Li, *Joint channel assignment and routing for throughput optimization in multi-radio wireless mesh networks*, in *Proceedings of the 11th Annual International Conference on Mobile Computing and Networking*, MobiCom '05, pp. 58–72, ACM, 2005.

[31] K. Jain, J. Padhye, V. N. Padmanabhan, and L. Qiu, *Impact of interference on multi-hop wireless network performance*, *Wireless networks* **11** (2005), no. 4 471–487.

[32] R. L. Cruz and A. V. Santhanam, *Optimal routing, link scheduling and power control in multihop wireless networks*, in *IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No. 03CH37428)*, vol. 1, pp. 702–711, IEEE, 2003.

[33] J. García-Rois, F. Gómez-Cuba, M. R. Akdeniz, F. J. González-Castaño, J. C. Burguillo, S. Rangan, and B. Lorenzo, *On the analysis of scheduling in dynamic duplex multihop mmWave cellular systems*, *IEEE Transactions on Wireless Communications* **14** (2015), no. 11 6028–6042.

[34] B. Li, D. Zhu, and P. Liang, *Small cell in-band wireless backhaul in massive multiple-input multiple-output systems*, in *Communications (ICC), 2015 IEEE International Conference on*, pp. 1838–1844, IEEE, 2015.

[35] S. W. Peters and R. W. Heath Jr, *The future of WiMAX: Multihop relaying with IEEE 802.16 j*, *IEEE Communications Magazine* **47** (2009), no. 1 104–111.

[36] Y. Li, J. Luo, W. Xu, N. Vucic, E. Pateromichelakis, and G. Caire, *A joint scheduling and resource allocation scheme for millimeter wave heterogeneous networks*, in *Wireless Communications and Networking Conference (WCNC), 2017 IEEE*, pp. 1–6, IEEE, 2017.

[37] E. Karamad, R. S. Adve, Y. Lostanlen, F. Letourneux, and S. Guivarch, *Optimizing placements of backhaul hubs and orientations of antennas in small cell networks*, in *Communication Workshop (ICCW), 2015 IEEE International Conference on*, pp. 68–73, IEEE, 2015.

[38] J. Du, E. Onaran, D. Chizhik, S. Venkatesan, and R. A. Valenzuela, *Gbps user rates using mmWave relayed backhaul with high-gain antennas*, *IEEE Journal on Selected Areas in Communications* **35** (2017), no. 6 1363–1372.

[39] R. Nelson and L. Kleinrock, *Spatial tdma: A collision-free multihop channel access protocol*, *IEEE Transactions on communications* **33** (1985), no. 9 934–944.

[40] A. Ephremides and T. V. Truong, *Scheduling broadcasts in multihop radio networks*, *IEEE Transactions on communications* **38** (1990), no. 4 456–460.

[41] S. A. Borbash and A. Ephremides, *Wireless link scheduling with power control and SINR constraints*, *IEEE Transactions on Information Theory* **52** (2006), no. 11 5106–5111.

[42] M. Johansson and L. Xiao, *Cross-layer optimization of wireless networks using nonlinear column generation*, *IEEE Transactions on Wireless Communications* **5** (2006), no. 2 435–445.

[43] S. Kompella, J. E. Wieselthier, A. Ephremides, H. D. Sherali, and G. D. Nguyen, *On optimal SINR-based scheduling in multihop wireless networks*, *IEEE/ACM Transactions on Networking (TON)* **18** (2010), no. 6 1713–1724.

[44] A. Capone, I. Filippini, and F. Martignon, *Joint routing and scheduling optimization in wireless mesh networks with directional antennas*, in *2008 IEEE International Conference on Communications*, pp. 2951–2957, IEEE, 2008.

[45] B. Hajek and G. Sasaki, *Link scheduling in polynomial time*, *IEEE transactions on Information Theory* **34** (1988), no. 5 910–917.

[46] C.-J. Lin, S.-H. Lin, and C.-F. Chou, *Performance study of optimal routing and channel assignment in wireless mesh networks*, in *IEEE GLOBECOM 2007-IEEE Global Telecommunications Conference*, pp. 4818–4822, IEEE, 2007.

[47] J. Luo, C. Rosenberg, and A. Girard, *Engineering wireless mesh networks: joint scheduling, routing, power control, and rate adaptation*, *IEEE/ACM Transactions on Networking (TON)* **18** (2010), no. 5 1387–1400.

[48] M. N. Kulkarni, A. Ghosh, and J. G. Andrews, *Max-min rates in self-backhauled millimeter wave cellular networks*, *arXiv preprint arXiv:1805.01040* (2018).

[49] B. Zhuang, D. Guo, and M. L. Honig, *Traffic-driven spectrum allocation in heterogeneous networks*, *IEEE J. Sel. Areas Commun. Special Issue on Recent Advances in Heterogeneous Cellular Networks* **33** (2015), no. 10 2027–2038.

[50] Z. Zhou and D. Guo, *1000-cell global spectrum management*, in *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, p. 20, ACM, 2017.

[51] Q. Kuang, W. Utschick, and A. Dotzler, *Optimal joint user association and multi-pattern resource allocation in heterogeneous networks.*, *IEEE Trans. Signal Processing* **64** (2016), no. 13 3388–3401.

[52] B. Zhuang, D. Guo, and M. L. Honig, *Energy-efficient cell activation, user association, and spectrum allocation in heterogeneous networks*, *IEEE Journal on Selected Areas in Communications* **34** (2016), no. 4 823–831.

[53] B. Zhuang, D. Guo, E. Wei, and M. L. Honig, *Large-scale spectrum allocation for cellular networks via sparse optimization*, *IEEE Transactions on Signal Processing* **66** (2018), no. 20 5470–5483.

[54] M. Pióro and D. Medhi, *Routing, flow, and capacity design in communication and computer networks*. Elsevier, 2004.

[55] O. Abari, H. Hassanieh, M. Rodriguez, and D. Katabi, *Millimeter wave communications: From point-to-point links to agile network connections.*, in *HotNets*, pp. 169–175, 2016.

[56] S. Bahmani and J. Romberg, *Efficient compressive phase retrieval with constrained sensing vectors*, in *Advances in Neural Information Processing Systems*, pp. 523–531, 2015.

[57] C. Rusu, R. Mendez-Rial, N. Gonzalez-Prelcic, and R. W. Heath, *Adaptive one-bit compressive sensing with application to low-precision receivers at mmwave*, in *2015 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, Dec, 2015.

[58] P. T. Boufounos and R. G. Baraniuk, *1-bit compressive sensing*, in *Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on*, pp. 16–21, IEEE, 2008.

[59] P. Schniter and S. Rangan, *Compressive phase retrieval via generalized approximate message passing*, *IEEE Transactions on Signal Processing* **63** (2015), no. 4 1043–1055.

[60] R. Pedarsani, D. Yin, K. Lee, and K. Ramchandran, *Phasecode: Fast and efficient compressive phase retrieval based on sparse-graph codes*, *IEEE Transactions on Information Theory* (2017).

[61] H. Ohlsson, A. Yang, R. Dong, and S. Sastry, *CPRL–an extension of compressive sensing to the phase retrieval problem*, in *Advances in Neural Information Processing Systems*, pp. 1367–1375, 2012.

[62] "IEEE 802.11 Task Group AD." `http://www.ieee802.org/11/Reports/tgad_update.htm`.

[63] B. Mamandipoor, D. Ramasamy, and U. Madhow, *Newtonized orthogonal matching pursuit: Frequency estimation over the continuum*, *IEEE Transactions on Signal Processing* **64** (2016), no. 19 5066–5081.

[64] E. J. Candes, X. Li, and M. Soltanolkotabi, *Phase retrieval via Wirtinger flow: Theory and algorithms*, *IEEE Transactions on Information Theory* **61** (2015), no. 4 1985–2007.

[65] E. J. Candes, *The restricted isometry property and its implications for compressed sensing*, *Comptes Rendus Mathematique* **346** (2008), no. 9 589–592.

[66] D. Ramasamy, S. Venkateswaran, and U. Madhow, *Compressive parameter estimation in AWGN*, *IEEE Trans. Signal Processing* **62** (2014), no. 8 2012–2027.

[67] M. Abdelghany, A. A. Farid, U. Madhow, and M. J. Rodwell, *Towards all-digital mmwave massive MIMO: Designing around nonlinearities*, in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, pp. 1552–1557, IEEE, 2018.

[68] S. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson, and C. Studer, *Throughput analysis of massive MIMO uplink with low-resolution ADCs*, *IEEE Transactions on Wireless Communications* **16** (2017), no. 6 4038–4051.

[69] E. Björnson, J. Hoydis, M. Kountouris, and M. Debbah, *Massive MIMO systems with non-ideal hardware: Energy efficiency, estimation, and capacity limits*, *IEEE Transactions on Information Theory* **60** (2014), no. 11 7112–7139.

[70] N. N. Moghadam, G. Fodor, M. Bengtsson, and D. J. Love, *On the energy efficiency of MIMO hybrid beamforming for millimeter-wave systems with nonlinear power amplifiers*, *IEEE Transactions on Wireless Communications* **17** (2018), no. 11 7208–7221.

[71] M. Wu, D. Wübben, A. Dekorsy, P. Baracca, V. Braun, and H. Halbauer, *Hardware impairments in millimeter wave communications using OFDM and SC-FDE*, in *WSA 2016; 20th International ITG Workshop on Smart Antennas*, pp. 1–8, VDE, 2016.

[72] D. Leeson, *A simple model of feedback oscillator noise spectrum*, *proc. IEEE* **54** (1966), no. 2 329–330.

[73] B. Razavi, *A study of phase noise in CMOS oscillators*, *IEEE journal of Solid-State circuits* **31** (1996), no. 3 331–343.

[74] A. Hajimiri and T. H. Lee, *A general theory of phase noise in electrical oscillators*, *IEEE journal of solid-state circuits* **33** (1998), no. 2 179–194.

[75] A. Mehrotra, *Noise analysis of phase-locked loops*, in *Proceedings of the 2000 IEEE/ACM international conference on Computer-aided design*, pp. 277–282, IEEE Press, 2000.

[76] Y. W. Kim and J. Du Yu, *Phase noise model of single loop frequency synthesizer*, *IEEE Transactions on Broadcasting* **54** (2008), no. 1 112–119.

[77] F. Herzel, S. A. Osmany, and J. C. Scheytt, *Analytical phase-noise modeling and charge pump optimization for fractional-N PLLs*, *IEEE Transactions on Circuits and Systems I: Regular Papers* **57** (2010), no. 8 1914–1924.

[78] S. Kalia, M. Elbadry, B. Sadhu, S. Patnaik, J. Qiu, and R. Harjani, *A simple, unified phase noise model for injection-locked oscillators*, in *2011 IEEE Radio Frequency Integrated Circuits Symposium*, pp. 1–4, IEEE, 2011.

[79] E. Björnson, M. Matthaiou, and M. Debbah, *Massive MIMO with non-ideal arbitrary arrays: Hardware scaling laws and circuit-aware design*, *IEEE Transactions on Wireless Communications* **14** (2015), no. 8 4353–4368.

[80] H. Thomas, V. Ranki, *et. al.*, *Phase noise in beamforming*, *IEEE Transactions on Wireless Communications* **9** (2010), no. 12 3682–3689.

[81] R. Combes and S. Yang, *An approximate ML detector for MIMO channels corrupted by phase noise*, *IEEE Transactions on Communications* **66** (2017), no. 3 1176–1189.

[82] Y.-F. Wang and J.-H. Lee, *A ZF-based precoding scheme with phase noise suppression for massive MIMO downlink systems*, *IEEE Transactions on vehicular Technology* **67** (2017), no. 2 1158–1173.

[83] Y.-F. Wang and J.-H. Lee, *A simple phase noise suppression scheme for massive MIMO uplink systems*, *IEEE Transactions on Vehicular Technology* **66** (2016), no. 6 4769–4780.

[84] A. Pitarokoilis, S. K. Mohammed, and E. G. Larsson, *Uplink performance of time-reversal MRC in massive MIMO systems subject to phase noise*, *IEEE Transactions on Wireless Communications* **14** (2014), no. 2 711–723.

[85] R. Krishnan, M. R. Khanzadi, N. Krishnan, Y. Wu, A. G. i Amat, T. Eriksson, and R. Schober, *Linear massive MIMO precoders in the presence of phase noisea large-scale analysis*, *IEEE Transactions on Vehicular Technology* **65** (2015), no. 5 3057–3071.

[86] A. Puglielli, G. LaCaille, A. M. Niknejad, G. Wright, B. Nikolić, and E. Alon, *Phase noise scaling and tracking in OFDM multi-user beamforming arrays*, in *2016 IEEE International Conference on Communications (ICC)*, pp. 1–6, IEEE, 2016.

[87] T. C. Schenk, X.-J. Tao, P. F. Smulders, and E. R. Fledderus, *On the influence of phase noise induced ICI in MIMO OFDM systems*, *IEEE Communications Letters* **9** (2005), no. 8 682–684.

[88] D. Petrovic, W. Rave, and G. Fettweis, *Effects of phase noise on OFDM systems with and without PLL: Characterization and compensation*, *IEEE Transactions on communications* **55** (2007), no. 8 1607–1616.

[89] S. Wu and Y. Bar-Ness, *OFDM systems in the presence of phase noise: consequences and solutions*, *IEEE Transactions on Communications* **52** (2004), no. 11 1988–1996.

[90] P. Robertson and S. Kaiser, *Analysis of the effects of phase-noise in orthogonal frequency division multiplex (OFDM) systems*, in *Proceedings IEEE International Conference on Communications ICC'95*, vol. 3, pp. 1652–1657, IEEE, 1995.

[91] M. H. Madani, A. Abdipour, and A. Mohammadi, *Analysis of performance degradation due to non-linearity and phase noise in orthogonal frequency division multiplexing systems*, *IET communications* **4** (2010), no. 10 1226–1237.

[92] K. Sathananthan and C. Tellambura, *Performance analysis of an OFDM system with carrier frequency offset and phase noise*, in *IEEE 54th Vehicular Technology Conference. VTC Fall 2001. Proceedings (Cat. No. 01CH37211)*, vol. 4, pp. 2329–2332, IEEE, 2001.

[93] L. Tomba, *On the effect of wiener phase noise in OFDM systems*, *IEEE Transactions on communications* **46** (1998), no. 5 580–583.

[94] T. Pollet, M. Van Bladel, and M. Moeneclaey, *Ber sensitivity of OFDM systems to carrier frequency offset and wiener phase noise*, *IEEE Transactions on communications* **43** (1995), no. 2/3/4 191–193.

[95] A. G. Armada and M. Calvo, *Phase noise and sub-carrier spacing effects on the performance of an OFDM communication system*, *IEEE Communications Letters* **2** (1998), no. 1 11–13.

[96] H. V. Poor and S. Verdú, *Probability of error in MMSE multiuser detection*, *IEEE transactions on Information theory* **43** (1997), no. 3 858–871.

[97] O. Momeni and E. Afshari, *High power terahertz and millimeter-wave oscillator design: A systematic approach*, *IEEE Journal of Solid-State Circuits* **46** (2011), no. 3 583–597.